Predicting Credit Status Case Study

Nengfeng Lin (Presentation, model, analysis), Chen Wang(Presentation, model, background) Yue Han(Presentation, model, conclusion), Xuankai Zhang(Presentation, model, explanation)

Introduction

Objective: build a model to predict the status of credit (good or bad) Interest: what factors lead to good or bad credit status and how can we use this to predict the status? In order to answer this question, we need to find a model that fits the data well.

Background and Significance

The data in this study is a stratified sample of 1000 credits (300 bad ones and 700 good ones) from the years 1973 to 1975 from a large regional bank in southern Germany. Although realistically, only 5% of data are bad ones, the bad credits in the dataset was clearly oversampled. Within the 20 explanatory variables of the dataset, seven were quantitative and 13 categorical. The importance of credit is well understood by all who live in this modern age, as it dramatically affects the capability to apply for a financial loan. Customers with “good” credit would compile with the contract terms, while those that are “bad” would not. Therefore, the bank would benefit if they could predict the status of credit with customers before the contract gave their information, which leads to this study. This analysis aims to build a model to predict the level of credit.

Load any libraries that will be used

```r
library(plyr)
library(tidyverse)
library(MASS)
library(ResourceSelection)
library(pROC)
```

Load the data set we will use to build the model

```r
url <- "Credit.csv"
credit<- read_csv(url, show_col_types = FALSE)
attach(credit)

## The following object is masked from package:MASS:
##
##     housing
```

Clean errors for data

We determine credit_risk to be our response variable since its description fits our objective. It tells us whether they have good or bad status of credit. By looking into the credit_risk variable, we see that in total we have 300 bad credits (who did not comply with the contract) and 700 good credits (who did comply with the contract)

```
credit %>% count(credit_risk)

## # A tibble: 2 × 2
##   credit_risk     n
##         <dbl> <int>
## 1           0   300
## 2           1   700
```

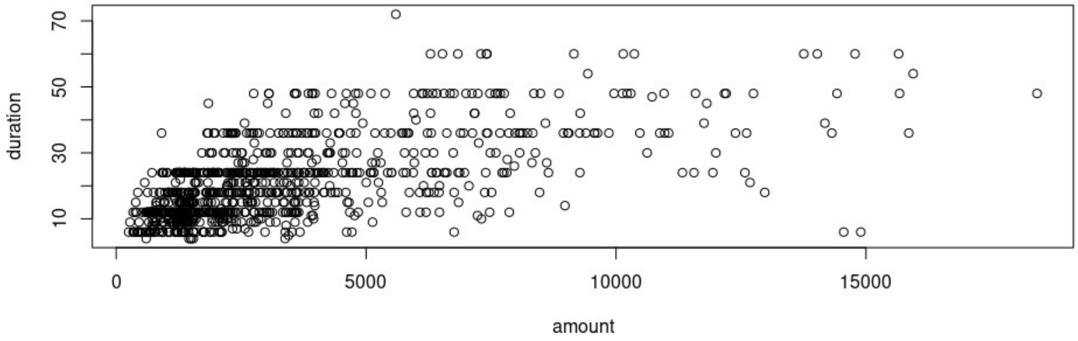
Establishing and classifying the variables, Quantitative or Qualitative

```
good <- as.factor(credit_risk==1)
bad <- as.factor(credit_risk==0)
creditHistory <- as.factor(credit_history)
status <- as.factor(status)
duration <- as.numeric(duration)
purpose <- as.factor(purpose)
amount <- as.numeric(amount)
savings <- as.factor(savings)
employDuration <- as.factor(employment_duration)
installmentRate <- as.factor(installment_rate)
personstatusSex <- as.factor(personal_status_sex)
otherDebtors <- as.factor(other_debtors)
presentResi <- as.factor(present_residence)
property <- as.factor(property)
age <- as.numeric(age)
othinsPlan <- as.factor(other_installment_plans)
housing <- as.factor(housing)
numCredit <- as.factor(number_credits)
job <- as.factor(job)
peoLiable <- as.factor(people_liable)
telephone <- as.factor(telephone)
forWorker <- as.factor(foreign_worker)
```

Exploratory Data Analysis

After loading the data, We need to check for the multicollinearity. The highest correlation value we get it 0.6 which is not very high so we do not need to worry about the correlation between any pair of the variables.

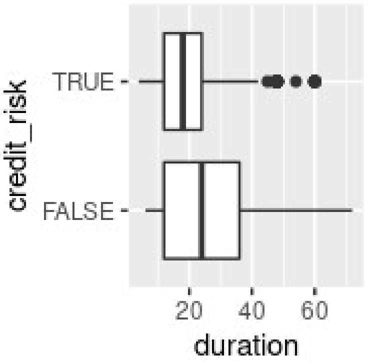
```
round(cor(amount,duration),2)

## [1] 0.62

plot(amount,duration)
```



We then decided to look at relationships within the data through plots Here are some plots of data we thought would relate to the status of credit (good or bad)

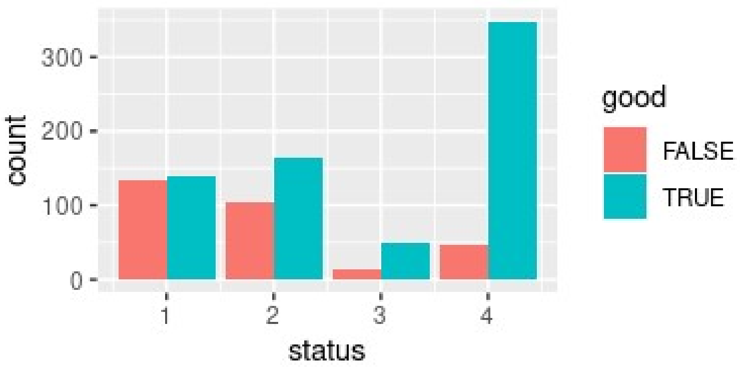
The first boxplot is duration and credit_risk (status of credit (good or bad)). We see there is a higher mode duration when there is bad status of credit which could lead us to believe there is a relationship between duration and credit status being good or bad

```
ggplot(aes(x = duration, y = good), data = credit) +geom_boxplot() +
ylab("credit_risk")
```

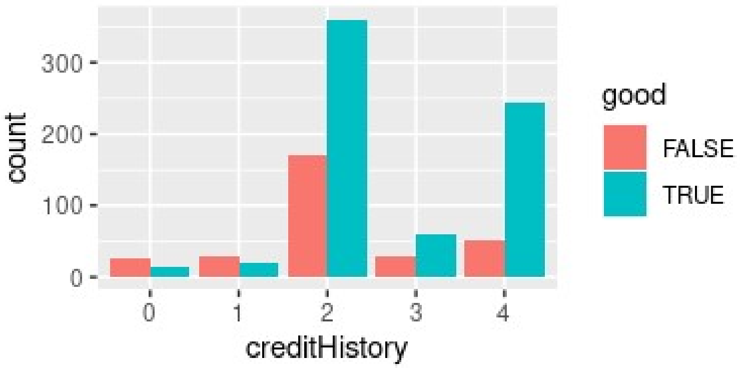


Here are some more graphs which shows certain status and creditHistory is related to credit status being good or bad such as when status is 4 (>= 200 DM / salary for at least 1 year) or when creditHistory = 2 (no credits taken/all credits paid back duly)

```
par( mfrow = c(2,2) )
ggplot(data = credit, aes(x = status, fill = good)) +
    geom_bar(position = "dodge")
```



```
ggplot(data = credit, aes(x = creditHistory, fill = good)) +
    geom_bar(position = "dodge")
```



## Model

The first model that we come up is the model with all variables, but no interation between any variable.

```
credit.fit1 <-
glm(good~creditHistory+status+duration+purpose+amount+savings+employDuration+
installmentRate+personstatusSex+otherDebtors+presentResi+property+age+othinsP
lan+housing+numCredit+job+peoLiable+telephone+forWorker,family = binomial)
summary(credit.fit1)

##
## Call:
## glm(formula = good ~ creditHistory + status + duration + purpose +
##     amount + savings + employDuration + installmentRate + personstatusSex
+
##     otherDebtors + presentResi + property + age + othinsPlan +
##     housing + numCredit + job + peoLiable + telephone + forWorker,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7232  -0.6838   0.3786   0.6931   2.3268
##
## Coefficients:
```

```
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.889e-01  1.169e+00   0.418 0.675921
## creditHistory1     -1.852e-01  5.687e-01  -0.326 0.744723
## creditHistory2      6.038e-01  4.432e-01   1.362 0.173054
## creditHistory3      9.735e-01  4.789e-01   2.033 0.042085 *
## creditHistory4      1.512e+00  4.462e-01   3.389 0.000701 ***
## status2             3.783e-01  2.189e-01   1.728 0.083918 .
## status3             9.630e-01  3.703e-01   2.600 0.009315 **
## status4             1.752e+00  2.339e-01   7.491 6.85e-14 ***
## duration           -2.833e-02  9.404e-03  -3.012 0.002594 **
## purpose1            1.629e+00  3.769e-01   4.322 1.54e-05 ***
## purpose2            7.287e-01  2.629e-01   2.772 0.005570 **
## purpose3            8.507e-01  2.482e-01   3.427 0.000609 ***
## purpose4            4.994e-01  7.667e-01   0.651 0.514860
## purpose5            1.943e-01  5.541e-01   0.351 0.725836
## purpose6           -1.163e-01  3.962e-01  -0.293 0.769167
## purpose8            1.915e+00  1.170e+00   1.637 0.101633
## purpose9            6.524e-01  3.362e-01   1.940 0.052332 .
## purpose10           1.387e+00  7.794e-01   1.780 0.075103 .
## amount             -1.200e-04  4.489e-05  -2.672 0.007536 **
## savings2            3.730e-01  2.914e-01   1.280 0.200586
## savings3            3.580e-01  4.006e-01   0.894 0.371527
## savings4            1.422e+00  5.339e-01   2.663 0.007748 **
## savings5            9.648e-01  2.648e-01   3.643 0.000269 ***
## employDuration2    -1.038e-02  4.392e-01  -0.024 0.981141
## employDuration3     2.868e-01  4.195e-01   0.684 0.494123
## employDuration4     8.251e-01  4.580e-01   1.802 0.071615 .
## employDuration5     2.415e-01  4.214e-01   0.573 0.566551
## installmentRate2   -2.812e-01  3.085e-01  -0.912 0.361976
## installmentRate3   -6.389e-01  3.404e-01  -1.877 0.060498 .
## installmentRate4   -9.390e-01  3.033e-01  -3.096 0.001961 **
## personstatusSex2    2.699e-01  3.880e-01   0.696 0.486693
## personstatusSex3    8.215e-01  3.803e-01   2.160 0.030742 *
## personstatusSex4    3.713e-01  4.558e-01   0.815 0.415306
## otherDebtors2      -4.328e-01  4.129e-01  -1.048 0.294598
## otherDebtors3       9.194e-01  4.263e-01   2.156 0.031047 *
## presentResi2       -7.444e-01  2.999e-01  -2.482 0.013060 *
## presentResi3       -5.068e-01  3.359e-01  -1.509 0.131306
## presentResi4       -3.694e-01  3.031e-01  -1.219 0.222914
## property2          -2.654e-01  2.545e-01  -1.043 0.297120
## property3          -1.662e-01  2.371e-01  -0.701 0.483317
## property4          -7.113e-01  4.238e-01  -1.678 0.093276 .
## age                 1.258e-02  9.273e-03   1.357 0.174865
## othinsPlan2        -3.507e-02  4.264e-01  -0.082 0.934448
## othinsPlan3         4.040e-01  2.474e-01   1.633 0.102437
## housing2            4.742e-01  2.363e-01   2.007 0.044755 *
## housing3            6.108e-01  4.795e-01   1.274 0.202768
## numCredit2         -3.896e-01  2.465e-01  -1.580 0.113996
## numCredit3         -3.155e-01  5.995e-01  -0.526 0.598763
## numCredit4         -3.676e-01  1.075e+00  -0.342 0.732389
## job2              -5.306e-01  6.795e-01  -0.781 0.434870
## job3              -5.397e-01  6.553e-01  -0.824 0.410159
## job4              -4.318e-01  6.661e-01  -0.648 0.516873
## peoLiable2         2.607e-01  2.513e-01   1.038 0.299474
## telephone2         2.677e-01  2.025e-01   1.322 0.186227
## forWorker2        -1.399e+00  6.223e-01  -2.248 0.024594 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.73  on 999  degrees of freedom
## Residual deviance:  892.96  on 945  degrees of freedom
## AIC: 1003
##
## Number of Fisher Scoring iterations: 5
```

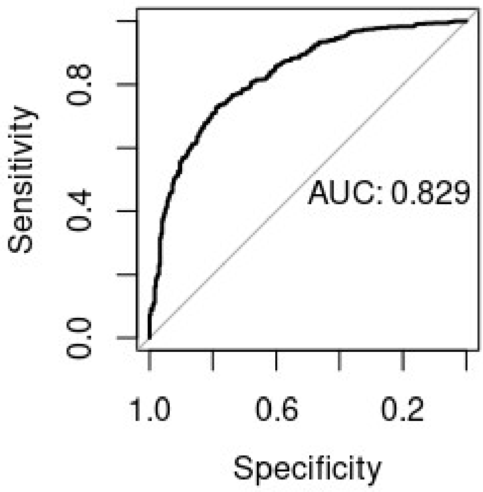
Model with no main effects and interactions

```
credit.fit2 <- glm(good~1,family = binomial)
```

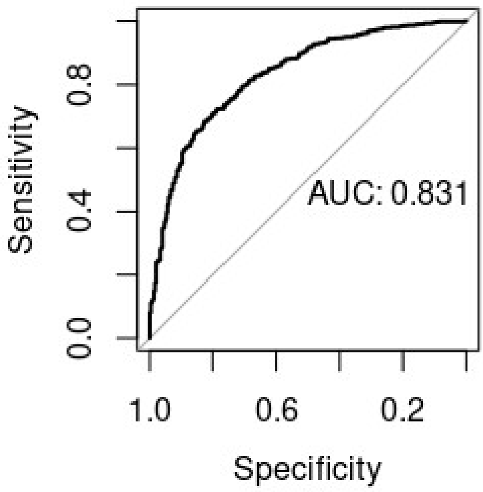
We see from the AIC value above, we need more variables Using stepwise, backward elimination, and forward selection we get the following

```
cred.both <- glm(formula = good ~ status + duration + creditHistory + purpose
+
    savings + otherDebtors + forWorker + presentResi + housing +
    installmentRate + amount + personstatusSex + telephone +
    othinsPlan, family = binomial)
#summary(cred.both)

#stepAIC(credit.fit1, direction="backward", scope=list(upper = credit.fit1,
lower=credit.fit2))

cred.backward <- glm(formula = good ~ creditHistory + status + duration +
purpose +
    amount + savings + employDuration + installmentRate + personstatusSex +
    otherDebtors + presentResi + age + othinsPlan + housing +
    forWorker, family = binomial)
#summary(cred.backward)

#stepAIC(credit.fit2, direction="forward", scope=list(upper = credit.fit1,
lower=credit.fit2))

cred.forward <- glm(formula = good ~ status + duration + creditHistory +
purpose +
    savings + otherDebtors + forWorker + employDuration + presentResi +
    housing + installmentRate + amount + personstatusSex + telephone +
    othinsPlan, family = binomial)
#summary(cred.forward)

credit_roc = roc(good~fitted(cred.both), plot=TRUE, print.auc = TRUE)
```

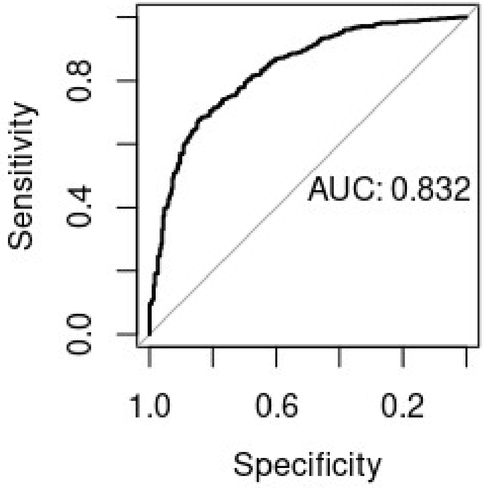
```
## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases
```



```
credit_roc_b = roc(good~fitted(cred.backward), plot=TRUE, print.auc = TRUE)

## Setting levels: control = FALSE, case = TRUE
## Setting direction: controls < cases
```



```
credit_roc_f = roc(good~fitted(cred.forward), plot=TRUE, print.auc = TRUE)

## Setting levels: control = FALSE, case = TRUE
## Setting direction: controls < cases
```



We use the forward selection model because it has the lowest AIC value and largest ROC value. The larger the concordance index the better. We see the forward model has the highest concordance index value c = 0.832

LRT indicates strong status, duration, creditHistory, purpose,savings, otherDebtors, forWorker, employDuration, presentResi, housing,installmentRate, amount and personstatusSex effect however telephone and othinsPlan does not indicate strong effect so maybe we can remove them to get a better model let's try this.

```
drop1(cred.forward, test = "Chisq")

## Single term deletions
##
## Model:
## good ~ status + duration + creditHistory + purpose + savings +
##     otherDebtors + forWorker + employDuration + presentResi +
##     housing + installmentRate + amount + personstatusSex + telephone +
##     othinsPlan
##                 Df Deviance     AIC    LRT  Pr(>Chi)
## <none>               902.14  990.14
## status           3   972.19 1054.19 70.056 4.152e-15 ***
## duration         1   912.94  998.94 10.804 0.0010128 **
## creditHistory    4   924.28 1004.28 22.142 0.0001878 ***
## purpose          9   936.17 1006.17 34.032 8.818e-05 ***
## savings          4   923.42 1003.42 21.282 0.0002785 ***
## otherDebtors     2   909.37  993.37  7.232 0.0268844 *
## forWorker        1   909.04  995.04  6.903 0.0086052 **
## employDuration   4   910.01  990.01  7.876 0.0962124 .
## presentResi      3   910.09  992.09  7.952 0.0470085 *
## housing          2   907.36  991.36  5.228 0.0732575 .
## installmentRate  3   915.64  997.64 13.502 0.0036680 **
## amount           1   910.07  996.07  7.937 0.0048441 **
## personstatusSex  3   910.34  992.34  8.205 0.0419526 *
## telephone        1   904.41  990.41  2.274 0.1315565
## othinsPlan       2   906.28  990.28  4.141 0.1261316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the model of forward selection with telephone and othinsPlan dropped we want to compare which one is better. We see AIC is slightly higher when dropping those variables

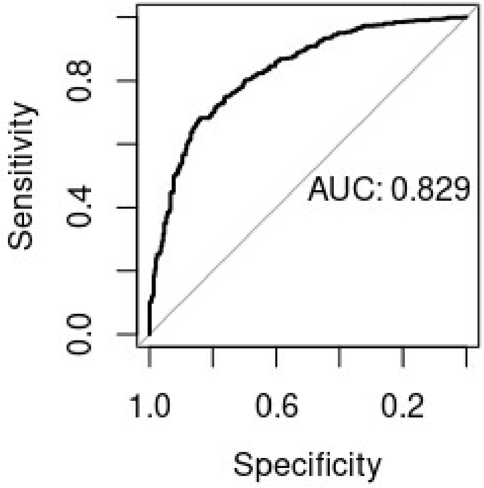
```
cred.fordrop <- glm(formula = good ~ status + duration + creditHistory +
purpose +
    savings + otherDebtors + forWorker + employDuration + presentResi +
    housing + installmentRate + amount + personstatusSex, family = binomial)
#summary(cred.fordrop)
```

since p-value <0.1 we can say the 90% anova test shows the forward model with dropped variables fits the data better but it is extremely close to 0.1 so we should look more into it

```
anova(cred.forward,cred.fordrop,test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: good ~ status + duration + creditHistory + purpose + savings +
##     otherDebtors + forWorker + employDuration + presentResi +
##     housing + installmentRate + amount + personstatusSex + telephone +
##     othinsPlan
## Model 2: good ~ status + duration + creditHistory + purpose + savings +
##     otherDebtors + forWorker + employDuration + presentResi +
##     housing + installmentRate + amount + personstatusSex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       956     902.14
## 2       959     908.43 -3  -6.2945  0.09813 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ROC The larger the concordance index the better. We see the forward model still has the highest concordance index value c = 0.832 from above

```
credit_roc_fd = roc(good~fitted(cred.fordrop), plot=TRUE, print.auc = TRUE)

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases
```



So because ROC and AIC are better in the forward selection model we will use it as our model.

sensitivity is 88% and specificity is 53% which is relatively good

```
n = dim(credit)[1]
prop = sum(credit$credit_risk==1)/n
prop2 = 0.5
y = (credit$credit_risk==1)*1
predicted = as.numeric(fitted(cred.forward) > prop2)
xtabs(~y + predicted)

##    predicted
## y     0   1
##   0 159 141
##   1  79 621

sensitivity = 621/(621+79)
sensitivity

## [1] 0.8871429

specificity = 159/(159+141)
specificity

## [1] 0.53
```

p-value > 0.05 fail to reject null hypothesis, the current model fits the data well

```
hoslem.test(cred.forward$y, fitted(cred.forward), g =16)

## 
##  Hosmer and Lemeshow goodness of fit (GOF) test
## 
## data:  cred.forward$y, fitted(cred.forward)
## X-squared = 13.594, df = 14, p-value = 0.4803
```

Conclusion

In conclusion we use the following model good ~ status + duration + creditHistory + purpose + savings + otherDebtors + forWorker + employDuration + presentResi + housing + installmentRate + amount + personstatusSex + telephone + othinsPlan

We can use this model to decide whether to approve a client for a contract based on our prediction This minimizes cases where clients do not comply with the contract

Some limitations we had was the small sized dataset which had bad credit oversampled and the data being old (sampled from 1973 to 1975). Since the ROC, AIC and anova test was so close it still can be up to discussion whether having a telephone landline and other installment plans is influential so we suggest doing further research on these two.

References

1. Grömping, U. (2019). Fachbereich II. FB II: Reports. Retrieved April 9, 2022, from http://www1.beuth-hochschule.de/FB_II/reports/welcome.htm

2. South German Credit (UPDATE) Data Set. UCI Machine Learning Repository: South German credit (update) data set. (n.d.). Retrieved April 1, 2022, from https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29