STAC51 Case Study Report

Under Instructor: Sohee Kang
10/04/2023

Factors that are Best Suited in Predicting Credit Risk of Individuals

Group 29

Primary Contributions

Jimmy Deng: Exploratory Data Analysis (#1006280472)
Yulun Wu: Modeling and Refinements (#1004912795)
Adil Shah: Conclusion and Discussion (#1004847151)
Raymond Chan: Editing and Background Research (#1004432269)

Libraries Used

```
library(tidyverse)
library(pander)
library(ggcorrplot)
library(corrr)
library(ResourceSelection)
library(pROC)
```

Introduction, Background and Significance

Credit plays a dynamic and pivotal role in all modern banking, shaping trust and economic futures (McLeay, 2014). Composed of both principal and interest, credit is the fundamental aspect of financial stability. Accurate ability to predict credit is key in running an economy with prosperity and fluidity. Such credit is used to provide people with the tools they need in their lives to build wealth, health, and create opportunity. Failure to allocate credit judiciously leads to unpaid debts, asset collapse, and ultimately failure of any modern banking system. Identifying key variables in predicting credit repayment is crucial for effective risk management and financial decision-making. In search for prediction, one may use a vast number of measures to predict and assess whether a debtor is likely to repay their debts. This case study finds and examines the most significant factors in predicting credit in a data set. It will be shown that status, duration, credit history, saving, age, and two interaction variables may strongly predict payment without the need for any additional data.

The primary objective of this case study is to offer valuable insights for assessing a debtor's trustworthiness, to aid lending decisions and minimizing financial risks.

Research Question: What factors are most important when predicting an individual's likelihood of repaying credit?

Exploratory Data Analysis

```r
credit_data <- read_csv("./credit.csv")     # Reading in the data set
```

Description of dataset

The data contains 1000 observations with 21 variables consisting of a mix of mostly categorical variables including binary, nominal, and ordinal variables, as well as a few continuous variables. The response variable is a binary variable called credit risk. The raw data set labels every variable with numbers and lacks meaningful descriptions, however, descriptions can be gotten through the provided 'CreditVariables-Description.R' script.

After cleaning and mutating data, the descriptions of the data set are as follows:

- Credit_risk (response), Ordinal; 0 = bad credit risk, 1 = good credit risk
- foreign_worker, Ordinal; 0 = not foreign worker, 1 = foreign worker
- telephone, Ordinal; 0 = no telephone #, 1 = has telephone # (under customer name)
- people_liable, Ordinal; 0 = one to two people, 1 = three or more people
- status, Ordinal; 1 = no acc., 2 = negative balance, 3 = balance >= 200 DM, 4 = 0 <= balance < 200 DM
- credit_history, Ordinal; 0 = delay in payments previously, 1 = critical acc./ existing credits, 2 = no credits taken/ fully paid on time, 3 = existing credits paid until now, 4 = fully paid on time
- purpose, Categorical; 0 = others, 1 = car (new), 2 = car (used), 3 = furniture/ equipment, 4 = radio/ television, 5 = domestic appliances, 6 = repairs, 8 = vacation, 9 = retraining, 10 = business
- savings, Ordinal; 1 = unknown/ no savings, 2 = balance < 100 DM, 3 = 100 <= balance < 500 DM, 4 = 500 <= balance < 1000 DM, 5 = balance >= 1000 DM
- personal_status_sex, Categorical; 1 = male: divorced/separated, 2 = female: non-single or male : single, 3 = male: married/widowed, 4 = female: single
- other_debtors, Ordinal; 1 = none, 2 = co-applicant, 3 = guarantor
- other_installment_plans, Categorical; 1 = bank, 2 = stores, 3 = none

- housing, Ordinal; 1 = for free, 2 = rent, 3 = own
- employment_duration, Ordinal/ discretized quantitative; 1 = unemployed, 2 = duration < 1 yr, 3 = 1 <= duration < 4 yrs, 4 = 4 <= duration < 7 yrs, 5 = duration >= 7 yrs
- installment_rate, Ordinal/ discretized quantitative; 1 = rate >= 35%, 2 = 25 <= rate < 35%, 3 = 20 <= rate <25%, 4 = < 20%
- present_residence, Ordinal/ discretized quantitative; 1 = duration < 1 yrs, 2 = 1 <= duration < 4 yrs, 3 = 4 <= duration < 7 yrs, 4 = duration >= 7 yrs
- property, Ordinal; 1 = unknown/ no property, 2 = car or other, 3 = building soc. savings agr./ life insurance, 4 = real estate
- number_credits, Ordinal; 1 = 1, 2 = 2-3, 3 = 4-5, 4 = >= 6
- job, Ordinal; 1 = unemployed/ unskilled (non-resident), 2 = unskilled (resident), 3 = skilled employee/ official, 4 = manager/ self empl./ highly qualified employee
- duration, Quantitative; Credit duration in months
- amount, Quantitative; Credit amount in DM
- age, Quantitative; Age of debtor in years

Cleaning and mutating data

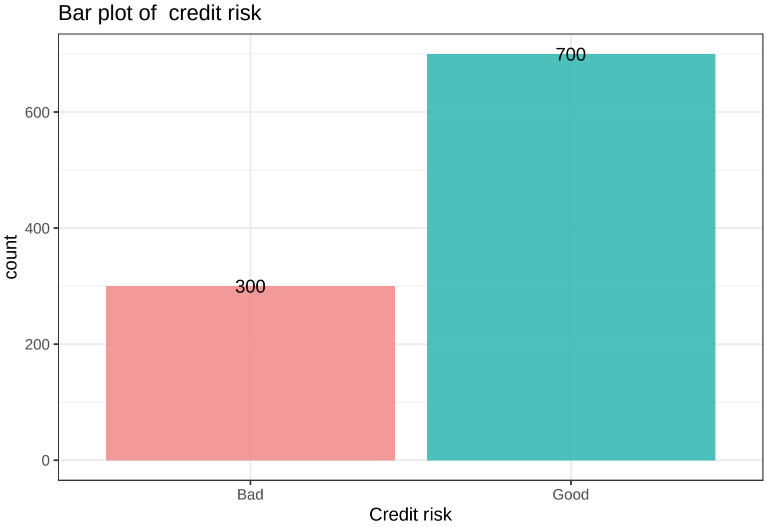
The data set came mostly clean, without any missing or invalid values. However, we changed some values: 1 - no telephone # => 0 - no telephone #, 2 - has telephone # => 1 - has telephone #, 2 - one to two people => 0 - one to two people to keep consistency, and to experiment with considering some variables as ordinal. Other than this, variables had to be factored appropriately since they were all numbers. We chose to not use the descriptions provided from the script as it was verbose. Also, we choose to treat some nominal variables as ordinal since their levels look like having some order relationship.

Visualizing variables

When looking at data, we decided to look at the distributions of the variables themselves, the distributions of the explanatory variables against response, and looking at correlation between continuous variables.

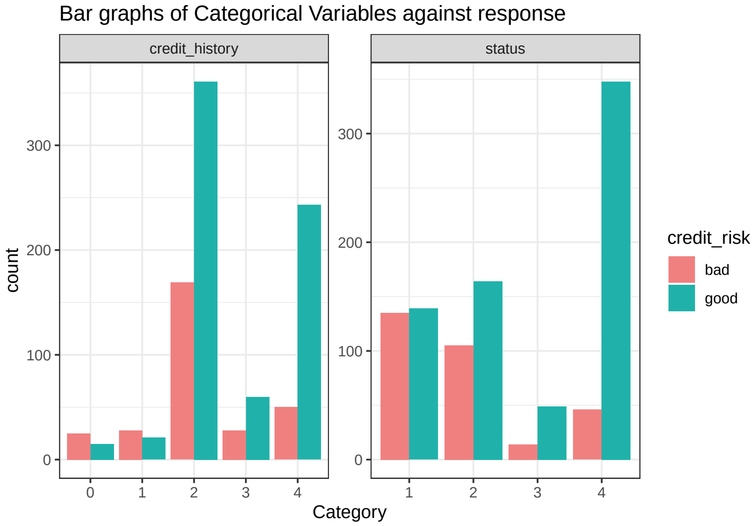
Distributions of variables

For most of the variables, nothing too notable stands out in the distributions, most variables end up skewed in some way. However, the distribution of the response variable credit_risk should be considered. Credit risk is skewed towards good risk, meaning that most proportions of bad risk will be low, which is important to keep in mind when comparing explanatory variables to the response variable.

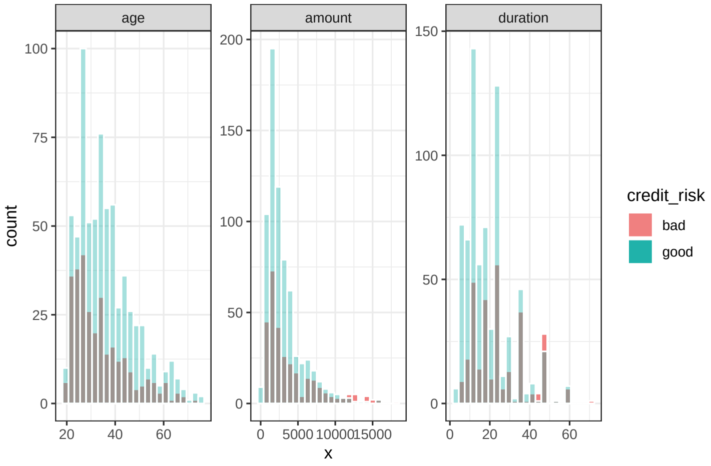


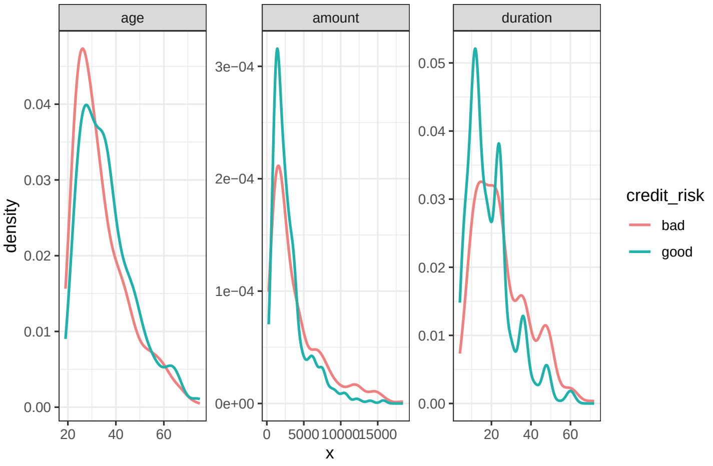
Distribution of explanatory variables against credit risk

As mentioned earlier, since observations are skewed towards good credit risk, most proportions of bad risk in categories will be low, so we should look for categories where the proportion is higher than usual. The explanatory variables credit_history and status are two notable examples of this. In categories 1 and 2 for these variables, the proportion of bad risk is close to equal of that for good risk, and even has a higher proportion in the case of credit_history. Given the high proportion of good risk in the data set, this suggests that these variables are more polarized and their categories have a clearer association to credit risk, and may help predict credit risk later on.



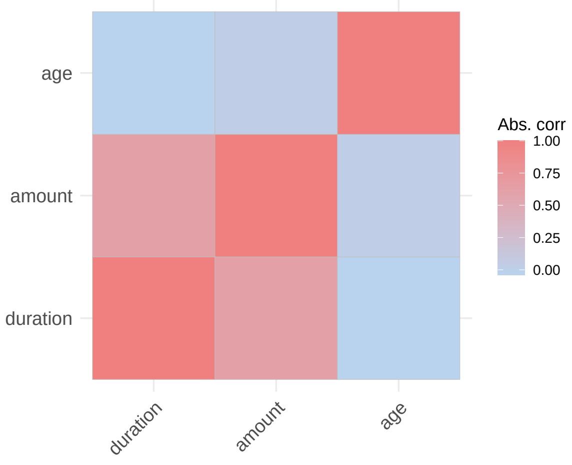
Comparing continuous variables to credit risk, it is easier to see where bad risk has higher proportions compared to good risk. For both amount and duration, bad risk become more prevalent as x increases, while for age, good risk becomes more prevalent as x increases.


Histograms of Continous Variables against response


Density plots of Continous Variables against response

Correlation between continuous variables

Lastly, we wanted to look at correlation between variables to see if there was any signs of multicolinearity. With techniques learned in this class, we could only compare correlation between continuous variables, so we did so. Between amount and duration, there appears to be significant correlation at 0.625, implying multicolinearity and as such, one of these variables may end up removed during the modelling process.


Correlation between continuous variables

	duration	amount	age
duration	1	0.625	-0.03755
amount	0.625	1	0.03227
age	-0.03755	0.03227	1

Model Building

```
# Separate data into training set and testing set, training set contains 700 observations, testing set
set.seed(1004912785)
credit_data.samp = base::sample(1:1000, size = 700, replace=FALSE)
credit_data.train = credit_data[credit_data.samp,]
credit_data.test = credit_data[-credit_data.samp,]
credit_data.main.best
```

```
##
## Call:  glm(formula = credit_risk ~ status + duration + credit_history +
##     purpose + amount + savings + employment_duration + installment_rate +
##     present_residence + property + age + housing + telephone,
##     family = binomial, data = credit_data.train)
##
## Coefficients:
##         (Intercept)               status             duration
##          -1.082e+00            5.507e-01           -2.243e-02
##      credit_history             purpose1             purpose2
##           2.615e-01            1.747e+00            7.155e-01
##            purpose3             purpose4             purpose5
##           7.493e-01            4.213e-01            7.038e-01
##            purpose6             purpose8             purpose9
##           3.823e-02            1.442e+01            1.688e-01
##           purpose10               amount              savings
##           2.067e+00           -9.431e-05            1.980e-01
## employment_duration     installment_rate    present_residence
##           1.601e-01           -2.632e-01           -1.923e-01
##            property                  age              housing
##          -3.047e-01            1.578e-02            3.772e-01
##           telephone
##           4.307e-01
##
## Degrees of Freedom: 699 Total (i.e. Null);  678 Residual
## Null Deviance:       843
## Residual Deviance: 662.7     AIC: 706.7
```

Forward, backward and stepwise eliminations give the same model, and its AIC is 706.7 which is smaller than full model (which gives 715.34). Thus, we select main effects: status, duration, credit_history, purpose, amount, savings, employment_duration, installment_rate, present_residence, property, age, housing, telephone.

Goodness-of-fit Test for Selected Main Effect Model

Since we have ungrouped data, so using Hosmer-Lemeshow test for Goodness-of-fit test.

```
hoslem.test(credit_data.main.best$y,fitted(credit_data.main.best),g=23)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  credit_data.main.best$y, fitted(credit_data.main.best)
## X-squared = 13.48, df = 21, p-value = 0.8909
```

Since p-value = 0.8909>0.05, we fail to reject the null hypothesis that the current model is as good as saturated model, thus this model fits data well, and consider the current model has less order than the saturated model, we choose this model and can try to add interaction terms to see if we can further improve the performance of our model.

Model With Interaction Term

Further select main effect

Since we still have too much parameters (22) after elimination, it will be hard to fit a saturated model with all of them, we will further select them to fit the saturated model. We want to divide these main effects into two categories and only use 4 of them with lowest AIC from finance category and 1-2 of them from personal info category.

```
# Finance models
model.status = glm(credit_risk~status,family=binomial,data=credit_data.train)
model.duration= glm(credit_risk~duration,family=binomial,data=credit_data.train)
model.credit_history= glm(credit_risk~credit_history,family=binomial,data=credit_data.train)
model.purpose= glm(credit_risk~purpose,family=binomial,data=credit_data.train)
model.amount= glm(credit_risk~amount,family=binomial,data=credit_data.train)
model.savings= glm(credit_risk~savings,family=binomial,data=credit_data.train)
model.installment_rate= glm(credit_risk~installment_rate,family=binomial,data=credit_data.train)
model.property= glm(credit_risk~property,family=binomial,data=credit_data.train)
model.installment_rate= glm(credit_risk~installment_rate,family=binomial,data=credit_data.train)
model.housing= glm(credit_risk~housing,family=binomial,data=credit_data.train)
```

```
# Personal info models
model.employment_duration= glm(credit_risk~employment_duration,family=binomial,data=credit_data.train)
model.present_residence= glm(credit_risk~present_residence,family=binomial,data=credit_data.train)
model.age= glm(credit_risk~age,family=binomial,data=credit_data.train)
model.telephone= glm(credit_risk~telephone,family=binomial,data=credit_data.train)
```

AIC of status is 761.21, AIC of duration is 821.4, AIC of credit_history is 821.35, AIC of purpose is 836.05, AIC of amount is 836.93, AIC of savings is 827.07, AIC of installment_rate is 843.28, AIC of property is 832.95, AIC of housing is 845.72.

4 of them with lowest AIC are: status, duration, credit_history, savings.

AIC of employment_duration is 840.53, AIC of present_residence is 845.78, AIC of age is 839.6, AIC of telephone is 844.05

4 of them with lowest AIC is: age.

```
hoslem.test(credit_data.select$y,fitted(credit_data.select),g=7)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  credit_data.select$y, fitted(credit_data.select)
## X-squared = 1.8491, df = 5, p-value = 0.8696
```

Since p-value = 0.8696>0.05, we fail to reject the null hypothesis that the current model is as good as saturated model, thus this model fits data well, and consider the current model has less order than the saturated model and less parameters than the model with 31 parameters, we choose this model and can try to add interaction terms to see if we can further improve the performance of our model.

```
credit_data.best
```

```
##
## Call:  glm(formula = credit_risk ~ status + duration + credit_history +
##     savings + age + duration:savings + credit_history:age, family = binomial,
##     data = credit_data.train)
##
## Coefficients:
```

```
##       (Intercept)             status            duration      credit_history
##           0.93182            0.57581            -0.06153            -0.37548
##           savings                age    duration:savings  credit_history:age
##          -0.18107           -0.02985             0.01575             0.01816
##
## Degrees of Freedom: 699 Total (i.e. Null);  692 Residual
## Null Deviance:       843
## Residual Deviance: 703.6     AIC: 719.6
```

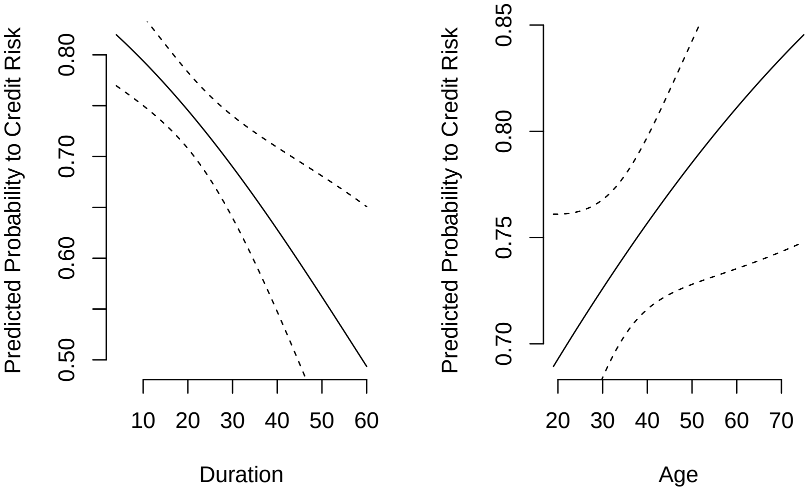
Both stepwise elimination and forward elimination gives the same model with AIC=719.6, but backward elimination gives saturated which has AIC=731.1. Although this model have slightly bigger AIC than the model the select using step function in previous section (the model with 22 parameters has AIC=706.7). We think its worth to sacrifice some AIC for about 1/2 less parameters. We select this model as our final model: $logit(y) = 0.93182 + 0.57581\beta_{status} - 0.06153\beta_{duration} - 0.37548\beta_{credit\_history} - 0.18107\beta_{savings} - 0.02985\beta_{age} + 0.01575\beta_{duration}\beta_{savings} + 0.01816\beta_{age}\beta_{credit\_history}$.

Test for Interaction Term

```
## Analysis of Deviance Table
##
## Model 1: credit_risk ~ status + duration + credit_history + savings +
##     age
## Model 2: credit_risk ~ status + duration + credit_history + savings +
##     age + duration:savings + credit_history:age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       694     716.49
## 2       692     703.59  2   12.903 0.001578 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value = 0.001578<0.05, we reject the null hypothesis that the simpler model (model without interaction) fits the data as good as model with interaction. Test statistic is 12.903, degrees of freedom is 2.

Predictive Probability Curve



From the predictive probability curve plotted based on the testing set on two continuous variables duration and age, one can see that as duration increases, the predictive probability of credit_risk decreases when holding other variables unchanged; as age increases, the predictive probability of credit_risk also increases when holding other variables unchanged.

Model Validation

Goodness-of-fit Test for Final Model

Since we have ungrouped data, so using Hosmer-Lemeshow test for Goodness-of-fit test.

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  credit_data.best$y, fitted(credit_data.best)
## X-squared = 2.885, df = 7, p-value = 0.8954
```

Since p-value = 0.8954>0.05, we fail to reject the null hypothesis that the current model is as good as saturated model, thus this model fits data well, and consider the current model has much less order than the saturated model, we choose this model as our final model.

Predictive Power

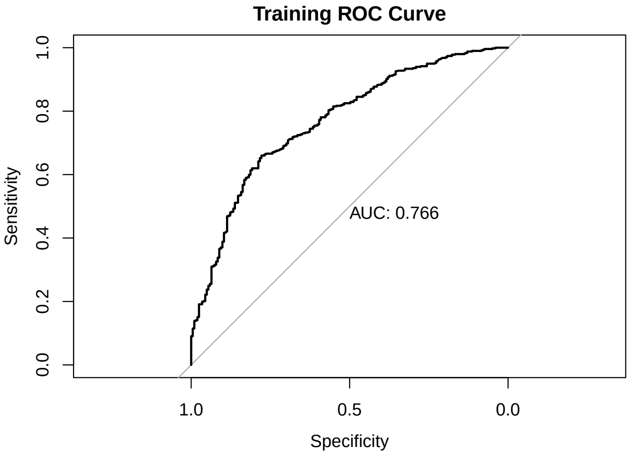
```
##                               predicted.train
## credit_data.train$credit_risk   0   1
##                              0 149  54
##                              1 163 334
```

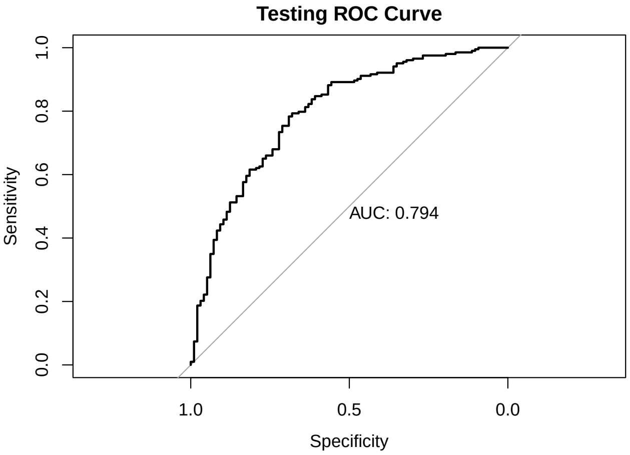
Sensitivity of final model on training set is 0.6720322. Specificity of final model on training set is 0.7339901. Concordance rate of final model on training set is 0.69.

```
##                              predicted.test
## credit_data.test$credit_risk   0   1
##                             0  64  33
##                             1  42 161
```

Sensitivity of final model on testing set is 0.7931034. Specificity of final model on testing set is 0.6597938. Concordance rate of final model on testing set is 0.75. From previous section, we see that the training Concordance rate is 0.69, this means our model generalize well on testing set.

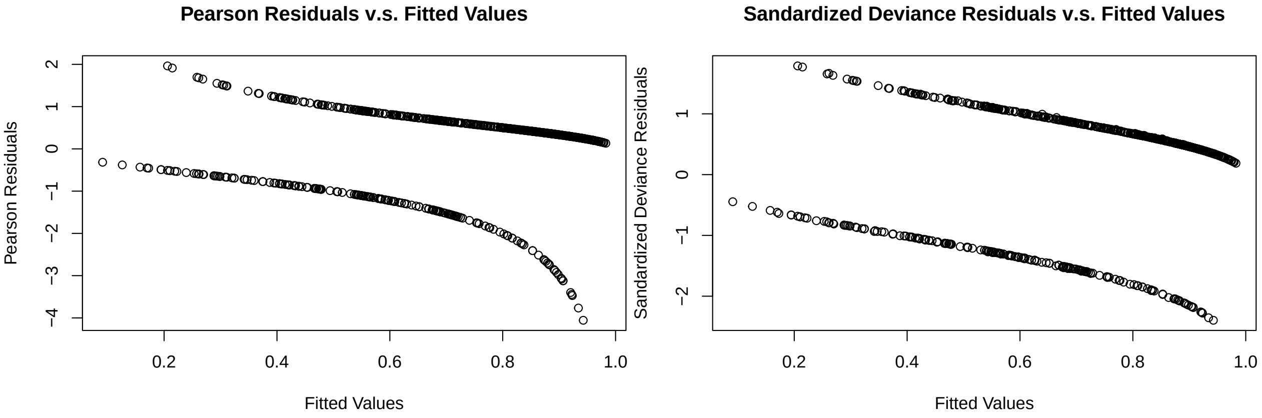
ROC





The area under training ROC curve is 0.766 and the area under testing ROC curve is 0.794, this means our model is good.

Residual Diagnosis



According to Pearson residual there are 9 outliers which has |Pearson residual| > 3 in training se. According to Standardized Deviance residual, seems like there is no outliers in training set (ie: no |standardized deviance residual| > 3).

Conclusion, Discussion, and Future Work

The goal of this report was to identify the factors that can affect someone's credit risk, and to predict the status of someone's credit based on those factors. In the final model, we concluded that the most important factors involved are status (the status of the debtor's checking account), duration (the duration of the credit), savings (debtor's savings), credit history (history of compliance with previous contracts), age and various interaction terms. One interesting fact we find is that if one's status gets worse by 1 level, the predicted probability of this person has good credit risk decreases by 57.6% when holding other variables unchanged, this means the amount of balance in one's account really impact credit risk a lot positively. And surprisingly, since status has such a big positive impact on credit risk, other main effects have a negative impact on credit risk as they increases, this is opposite to our common knowledge. Our finding could impact the field as it can help banks and other financial institutions in limiting credit risk, by helping them to assess which people have good or bad credit risk.

The major challenge we had was originally the model was fitted with Poisson, which gave a model with only status and duration as parameters with an AIC of 1315. However this model did not fit the data well, but due to an error with grouping numbers in the Hosmer-Lemeshow test, it seemed as if it did. After realizing this, we changed the model to binomial, which fit the data much better, and caused the AIC to decrease drastically.

Some limitations of the model are that the model we used to add interaction terms does not include all the significant variables that were found in the main effect model, so we might miss some important interaction terms. Another limitation is that we didn't select the model with the best AIC, as it had many more parameters, which could have led to other issues. Additionally the data is from 1970s Germany, meaning that comparing results derived from this data could be difficult due to differences in culture and the time period. Another issue is that some of the variables look ordinal, but are treated as nominal. For future research it would be interesting to treat these variables as ordinal, and examine how it affects the model, as well as collecting more recent data to determine how the factors affecting credit risk have changed over the last 50 years.

One future work we can do is can try to add interaction terms using the main effect model with all significant variables. This could take long time as the saturated model will be huge.

References

Grömping, U. (2019). South German credit data: Correcting a widely used data set (Report No. 4/2019). Beuth University of Applied Sciences Berlin. https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29

McLeay, M., Radia, A., & Thomas, R. (2014). Money creation in the modern economy. Bank of England Quarterly Bulletin, 2014 Q1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=24162341