STAC51 Case Study Report

Artem Petrishchev, Daniel Ekoko, Niranj Sasikumar, Ting Lei

Building a Model to Predict the Status of Credit

Group 30

Artem Petrishchev, 1002575260: Background and Significance, Exploratory Data Analysis
Ting Lei, 1005813425: Background and Significance, Exploratory Data Analysis
Daniel Ekoko, 1003551177: Model Selection, Model Validation/Diagnostics
Niranj Sasikumar, 1005070539: Model Selection, Model Validation/Diagnostics

Word Count: 1797

Background and Significance

Abstract

Credit risk refers to the risk of default on a loan or other credit extended to an individual or entity. It is a measure of the likelihood that a borrower will not be able to repay their debts as agreed. Credit risk is a significant concern for lenders, who must manage this risk in order to ensure that they are able to recover the funds they have loaned out. The background of credit risk associated with individuals dates back to the early days of lending. As people began to borrow money for various purposes, lenders realized that they needed to assess the credit-worthiness of borrowers in order to minimize the risk of default. Over time, this led to the development of credit scoring systems, which use various factors to assess the likelihood that a borrower will repay their debts.

The significance of credit risk associated with individuals is that it can have significant financial implications for both the borrower and the lender. For the borrower, a poor credit score or history of default can make it more difficult to obtain credit in the future, and may result in higher interest rates and fees. For the lender, credit risk can lead to losses if borrowers default on their loans, which can have a ripple effect on the lender's ability to continue lending.

This case study aims to to analyze the factors contributing to credit risk associated with individual borrowers, and to understand the relationship between certain personal and financial factors in determining credit risk.

Research Question: Can we predict credit risk for individuals using personal information and financial history?

Variable Description

- Status - Status of debtor's checking account (no checking account, <0 DM, 0<=...<200 DM, >=200 DM)
- Duration - Credit duration in months
- Credit History - History of compliance (delay in paying off in the past, critical account, no credits taken, existing credits paid back duly till now, all credits at this bank paid back duly)
- Purpose - Purpose for which the credit is for (others, car (new), car(used), furniture, radio/television, domestic appliances, repairs, education, vacation, retraining, business)
- Amount - Credit amount in DM (German currency)
- Savings - Debtors savings (no savings account, < 100 DM, 100 <= ... < 500 DM, 500 <= ... < 1000 DM, ... >= 1000 DM)
- Employment Duration - Duration of debtor's employment with current employer (unemployed, < 1yr, 1 <= ... < 4 yr, 4 <= ... < 7 yr, >= 7yr)
- Installment Rate - Credit installments as a percentage of debtor's disposable incomes (>= 35, 25 <= ... < 35, 20 <= ... < 25, < 20)
- Personal Status Sex - Sex and marital status (male: divorced/separated, female: non-single or male:single, male: married/widowed, female: single)
- Other Debtors - Another debtor (none, co-applicant, guarantor)
- Present Residence - Length of time in years that the debtor has lived in their present residence (< 1yr, 1 <= ... < 4 yr, 4 <= ... < 7 yr, >= 7 yr)
- Property - Debtors most valued property (unknown/no property, car or other, building soc. savings agr./life insurance, real estate)

- Age - Age in years
- Other Installment Plans - Installment plans from providers (bank, stores, none)
- Housing - Type of housing the debtor lives in (for free, rent, own)
- Job - Quality of debtor's job (unemployed/unskilled - non-resident, unskilled - resident, skilled employee/official, manager/self-empl./highly qualif. employee)
- People Liable - Number of persons who are financially dependent on the debtor (3 or more, 0 to 2)
- Telephone - Telephone land-line registered on the debtors name (no, yes (under customer name))
- Foreign Worker Status - Is debtor a foreign worker (no, yes)
- Credit Risk - Has the credit contract been complied with (good, bad)

Exploratory Data Analysis

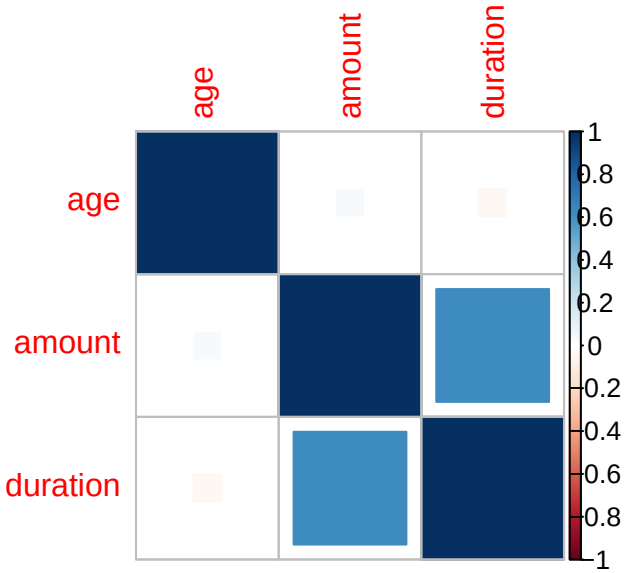
Loading the Data

```
data <- read.csv("data.updated.csv")
data <- data %>% mutate_if(is.character, as.factor)
```

We loaded the data and factored all qualitative variables in the data. The data set contains 1000 observations. Depending on the information we obtain from the data set, there are 20 predictors we consider that may have an effect on the status of credit.

Analysis of Quantitative Variables

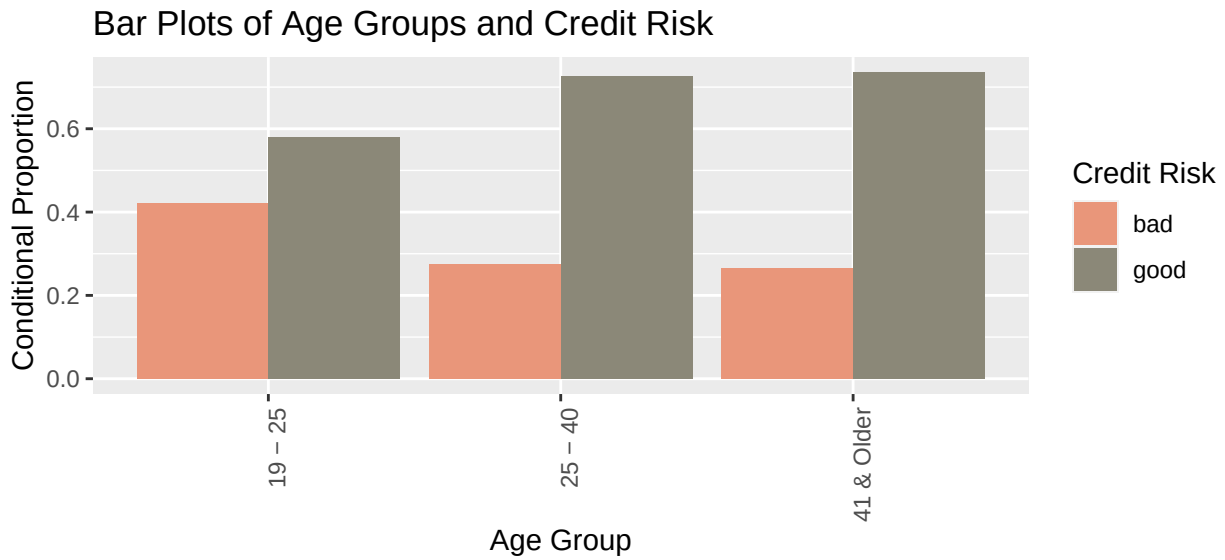
```
quantitative <- c("age", "amount", "duration")
df <- data[, quantitative]
corrplot(cor(df), method = "square")
```



First, a quantitative variable analysis was performed. We examined the correlation between the quantitative variables. We can see that there is a moderate correlation between duration and amount. However, there are no strongly correlated predictor variables.

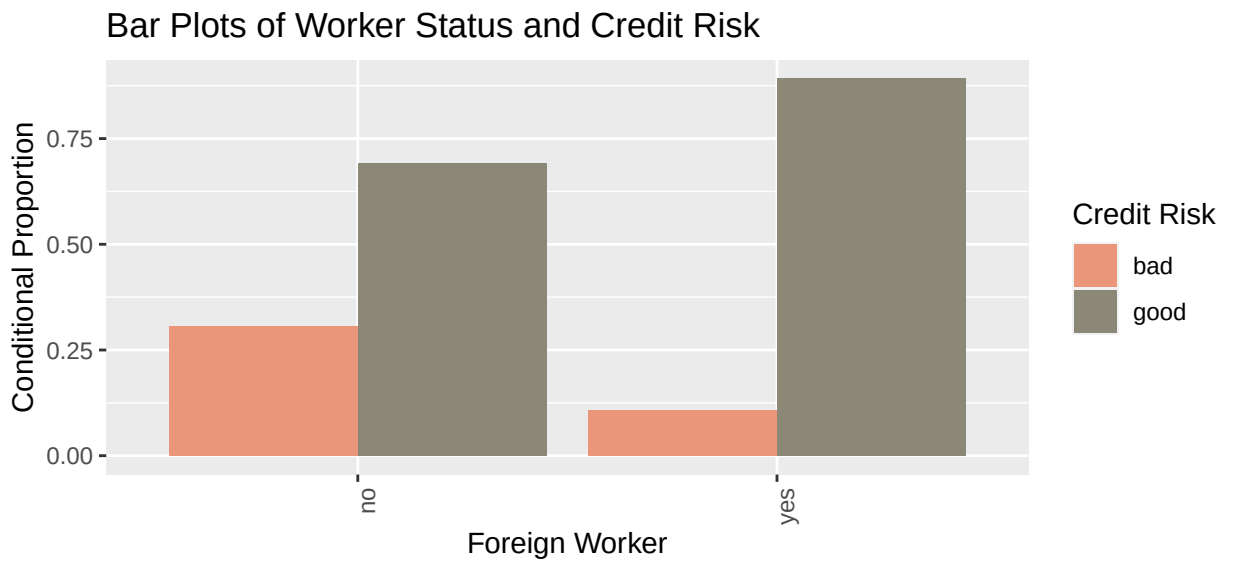
Analysis of Qualitative Variables

```
# Group by age ranges
data <- data %>% mutate(age.update = case_when(age <= 25 ~ "19 - 25",
                                               age <= 40 ~ "25 - 40",
                                               age > 40 ~ "41 & Older"))
denoms <- c(190, 190, 538, 538, 272, 272)

ggplot(data = data, aes(x = factor(age.update), fill = factor(credit_risk))) +
  geom_bar(position = "dodge", aes(y = (..count..)/denoms)) + labs(fill = "Credit Risk") +
  ggtitle("Bar Plots of Age Groups and Credit Risk") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + xlab("Age Group") +
  ylab("Conditional Proportion") + scale_fill_manual(values = c("darksalmon", "cornsilk4"))
```



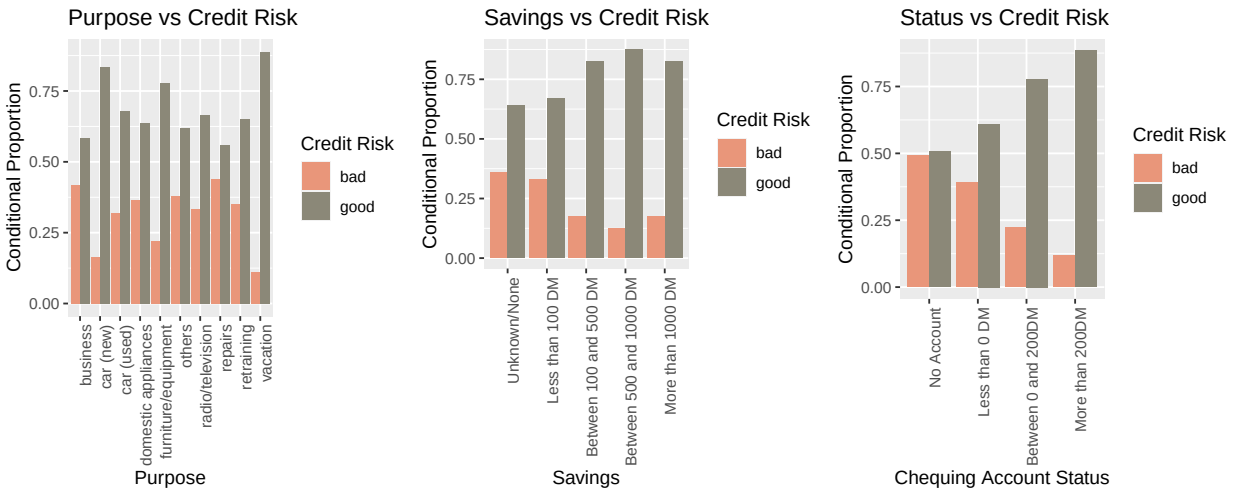
The plot suggests that there is a negative association between age and having "bad" credit risk. That is as age increases, credit risk decreases. The age group of 19 - 25 have the highest credit risk.

```
# We are dividing all counts by the total group count to get a proportionate plot view
denoms <- c(963, 963, 37, 37)

ggplot(data = data, aes(x = factor(foreign_worker), fill = factor(credit_risk))) +
  geom_bar(position = "dodge", aes(y = (..count..)/denoms)) + labs(fill = "Credit Risk") +
  ggtitle("Bar Plots of Worker Status and Credit Risk") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + xlab("Foreign Worker") +
  ylab("Conditional Proportion") + scale_fill_manual(values = c("darksalmon", "cornsilk4"))
```



The plot suggests that local workers have relatively higher credit risk than foreign workers, and this suggests that there is an association between worker status and credit risk. Another point worth mentioning is that we had considerably more local workers than foreign workers in the data.

```
grid.arrange(purpose_plot, savings_plot, status_plot, ncol = 3)
```



Analysis Summary

Based on the correlation matrix, no strongly correlated predictor variables were observed so multi-collinearity is not an issue and none of the variables need to be dropped. Secondly, we observed a negative association between age and credit risk. As age increases, credit risk decreases. Another relationship that we considered was that between foreign worker status and credit risk. Although our plot suggests that local workers have higher credit risk than foreign workers, we must note that there are considerably more local workers than foreign workers in our data.

Model Selection

Splitting the Data

```
set.seed(123)
train_index <- createDataPartition(data$credit_risk, p = 0.8, list = FALSE)
train_data <- data[train_index,]
test_data <- data[-train_index,]
```

We split the data into training data and testing data.

Automatic Model Selection

```
full_model <- glm((credit_risk == "bad") ~ ., data = train_data, family = "binomial")
```

We chose a logistic model since credit_risk is a binary variable. We started by fitting a main effect model with all predictors. We started with a model that had no interaction terms since there would be too many terms even if we only considered two-way interactions.

```
model_both <- step(full_model, direction = "both", trace = 0)
model_backward <- step(full_model, direction = "backward", trace = 0)
model_forward <- step(glm(credit_risk ~ 1, data = train_data, family = "binomial"),
                      direction = "forward", scope = formula(full_model), trace = 0)
```

We used the step() function to automatically select the model. We experimented with different directions.

```
formula(model_backward)
```

```
## (credit_risk == "bad") ~ status + duration + credit_history +
##     purpose + amount + savings + installment_rate + personal_status_sex +
##     other_debtors + present_residence + property + age + other_installment_plans +
##     housing + foreign_worker
```

All directions selected the same predictors with the same AIC and the same residual deviance.

We considered the two-way interactions between all the potential predictors in our data and found that this provided no additional value to the performance of our model based on the fact that adding any possible combination of two-way interactions resulted in higher AIC's for our model.

Manual Model Selection

```
manual_model = glm((credit_risk == "bad") ~ status + duration + credit_history +
                     purpose + amount + savings + installment_rate + personal_status_sex +
                     other_debtors + age + foreign_worker, family = "binomial",
                   data = train_data)
```

We used the drop1() function to try and further simplify the automatically-selected model by dropping the least significant variables using a significance level of 0.05. The variables that we dropped using this method were present_residence, property, other_installment_plans, and housing.

Comparing the Automatically-Selected and Manually-Selected Models

```
##           Automatic Model Manual Model
## AIC           778.1763610  781.3210876
## Accuracy        0.6850000    0.7100000
## Precision       0.4814815    0.5151515
## F-Score         0.5531915    0.5396825
## ROC AUC         0.8443824    0.8324256
```

Although the manually-selected model was simpler in terms of having less variables, the AIC of the model increased and the area under the ROC curve decreased. Although, both ROC AUC's suggest that both models are excellent fits. The accuracy and precision of the manually-selected model on the testing dataset increased, but the F-score decreased. So although the manually-selected model is worse in terms of AIC, ROC, and F-score, it has better accuracy and precision on the testing data.

```
anova(manual_model, automatic_model, test = "LRT")$"Pr(>Chi)"[2]
```

```
## [1] 0.01022479
```

The likelihood ratio test (LRT) is an appropriate goodness-of-fit test for our analysis because it is designed to compare nested models, which is the case for our two candidate models, as one is a reduced version of the other. Our LRT findings revealed a significant difference between the two models, with a p-value of 0.01022, which is less than the common significance level of 0.05. This result indicates that the more complex model, the automatically-selected model, provides a significantly better fit to the data compared to the simpler manually-selected model. Therefore, we will use the automatically-selected model for further analysis.

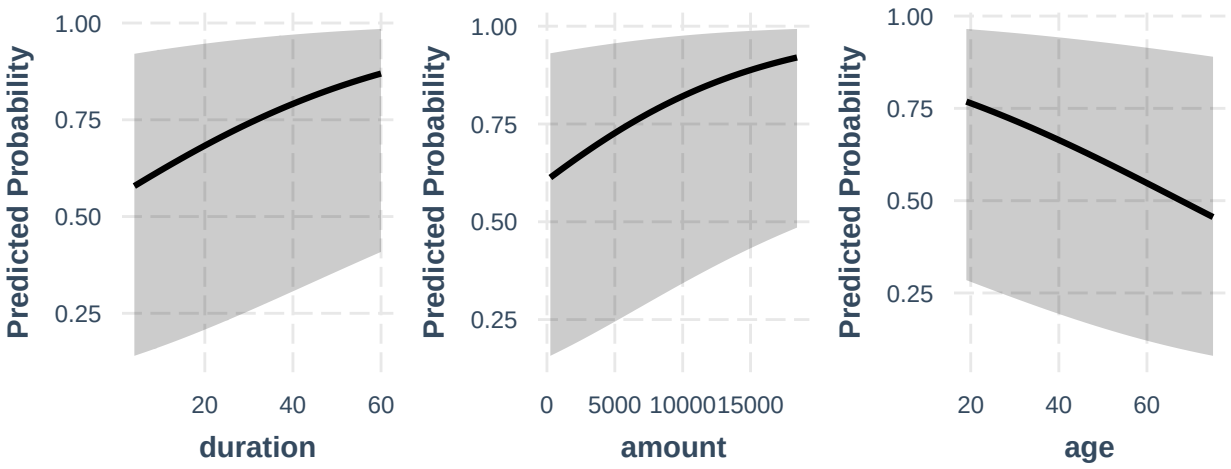
Model Validation/Diagnostics

```
hoslem.test(automatic_model$y, fitted(automatic_model), g = 17)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  automatic_model$y, fitted(automatic_model)
## X-squared = 16.293, df = 15, p-value = 0.3628
```

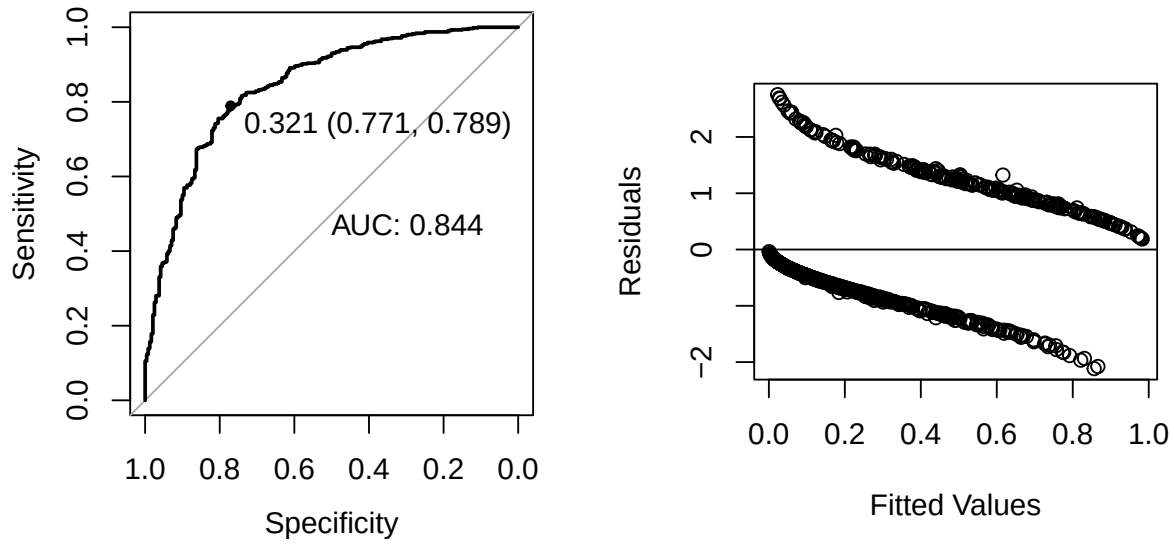
By running the Hosmer-Lemeshow test on the model, we get a p-value of 0.3628. This p-value is greater than our significance level of $\alpha = 0.05$, which means that our model fits the data well.

```
plot1 <- effect_plot(automatic_model, pred = duration, interval = TRUE, data = train_data,
                     y.label = "Predicted Probability")
plot2 <- effect_plot(automatic_model, pred = amount, interval = TRUE, data = train_data,
                     y.label = "Predicted Probability")
plot3 <- effect_plot(automatic_model, pred = age, interval = TRUE, data = train_data,
                     y.label = "Predicted Probability")
grid.arrange(plot1, plot2, plot3, ncol = 3)
```



From the predicted probability curves, we can see that the predicted probability of having bad credit status increases with credit duration and credit amount and decreases with age.

```
par(mfrow = c(1, 2))
roc <- roc(train_data$credit_risk ~ fitted(automatic_model), plot = TRUE,
           print.auc = TRUE, print.thres = "best")
residuals <- rstandard(automatic_model)
fitted <- fitted(automatic_model)
plot(fitted, residuals, xlab = "Fitted Values", ylab = "Residuals", abline(h = 0))
```



The area under the ROC curve is 0.844 which suggests that our model is a great fit. The best threshold or cutoff value for our predicted probabilities is 0.321. From the residual plot, we can conclude that there are no obvious outliers.

```
pred_prob <- predict(automatic_model, test_data, type = "response")
predicted <- as.factor(ifelse(pred_prob > 0.321, "bad", "good"))
```

```
confusion_matrix <- confusionMatrix(predicted, test_data$credit_risk)
confusion_matrix$table
```

```
##           Reference
## Prediction bad good
##       bad   39   42
##       good  21   98
```

```
confusion_matrix$byClass
```

```
##          Sensitivity          Specificity       Pos Pred Value
##            0.6500000            0.7000000            0.4814815
##       Neg Pred Value            Precision               Recall
##            0.8235294            0.4814815            0.6500000
##                   F1           Prevalence       Detection Rate
##            0.5531915            0.3000000            0.1950000
## Detection Prevalence    Balanced Accuracy
##            0.4050000            0.6750000
```

The sensitivity, specificity, and recall of our model on the testing data is fairly good. The precision is somewhat low compared to the other metrics, however, it is not an issue in our case since it is more important for a bank to correctly predict if the credit status will be "bad" rather than predicting if it will be "good".

Discussion/Conclusion

The goal of this report was to identify the factors that are likely to influence credit risk in individuals and build a model to predict whether the status of credit will be good or bad. From our analysis, we identified that the variables status, duration, credit_history, purpose, amount, savings, installment_rate, personal_status_sex, other_debtors, present_residence, property, age, other_installment_plans, housing, and foreign_worker are the best variables to have in our model to predict what credit_risk will be.

Analyzing factors that affect credit risk in individuals allows for significant research in this field and can advance the financial industry in various ways, such as improving credit assessment and underwriting processes.

Some of the limitations of our analysis would be the dataset being relatively small for such a task since it only contains 1000 observations and there were a lot of categorical variables which are more difficult to work with and interpret, this was in addition to many ordinal variables with more than two levels making it even more challenging to apply regression techniques effectively. Additionally, the dataset may contain bias or inaccurate representation and the dataset contains more samples of good credit than bad credit which may affect the classifier when we train it. Despite all of these limitations, we believe that our model has achieved the goal of identifying factors to be considered when assessing credit worthiness and predicting credit risk based on these factors.

For future research, it would be interesting to know by how much the factors we identified can increase or decrease credit risk. In other words, while we were able to identify factors that impact credit risk, there is still the question of which factors have the most significant influence on increasing or decreasing the credit risk in individuals. Furthermore, since the data is from the 1970s (Grömping, 2019), for future research, a possibility is to do the same analysis but with more recent data. There may be several new variables that need to be included. It may also be required to remove some of the variables that are used in this data such as the variable "telephone" since whether someone has a landline registered under their name may not be as helpful today as it was in the 1970s to predict credit risk.

References

Grömping, U. (2019). *South German Credit Data: Correcting a Widely Used Data Set* (Report No. 04/2019). Beuth University of Applied Sciences Berlin. Retrieved from http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf