Fachbereich II – Mathematik - Physik - Chemie

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

04/2019

Ulrike Grömping

South German Credit Data: Correcting a Widely Used Data Set

South German Credit Daten: Korrektur eines vielgenutzten Datensatzes (englischsprachig)

Reports in Mathematics, Physics and Chemistry

Berichte aus der Mathematik, Physik und Chemie

ISSN (print): 2190-3913

Reports in Mathematics, Physics and Chemistry

Berichte aus der Mathematik, Physik und Chemie

The reports are freely available via the Internet:
http://www1.beuth-hochschule.de/FB_II/reports/welcome.htm

04/2019, November 2019

South German Credit Data: Correcting a Widely Used Data Set
South German Credit Daten: Korrektur eines vielgenutzten Datensatzes
(englischsprachig)

Editorial notice / Impressum

Published by / Herausgeber:
Fachbereich II
Beuth Hochschule für Technik Berlin
Luxemburger Str. 10
D-13353 Berlin
Internet: http://public.beuth-hochschule.de/FB_II/
E-Mail: fbiireports@beuth-hochschule.de

ISSN (print): 2190-3913

South German Credit Data: Correcting a Widely Used Data Set

Ulrike Grömping

29 November 2019

1 Abstract

The widely used German credit dataset from the UCI Machine Learning Repository, donated by the German professor Hans Hofmann via the European Statlog project, comes with an incorrect code table. Many variables are wrongly represented, which implies that the data cannot be adequately used for experimenting with methods for interpretable machine learning. This note provides details on data generation and a correction of the code table, derived from German language literature. Correct data have been provided to the UCI Machine Learning Repository under the name "South German Credit" (https://archive.ics.uci.edu/ml/datasets/South+German+Credit).

2 Introduction

The UCI Machine Learning Repository (Dua and Graff 2019) contains a dataset on German credit data (https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data). It has been provided as part of a collection of datasets from an EU project called "Statlog" and will be called the Statlog German credit data in the following. The entry in the UCI Machine Learning Repository mentions prof. Hans Hofmann from Hamburg university (1994) as the donor of the dataset. The dataset has been widely used in machine learning research: for example, the Google search `"German credit" UCI Hofmann` yielded 4240 hits (27 November, 9am, Berlin, Germany). There are also several R packages that include these data: **evtree**, **CollapseLevels**, **caret**, **gamclass**, **klaR**, **rchallenge**, **scorecard** (checked on November 27, 2019, no guarantee for completeness). When I started to use these data for experimenting with methods for interpretable machine learning (IML), I soon realized that something must be wrong: for example, creditability of a debtor would become worse when changing the credit history variable to "no credits taken/all credits paid back duly" and better for changing the status variable to "no checking account" (as opposed to, e.g., an account at the credit-giving bank with a good balance). Further inspection showed that there are more implausibilities, e.g. that more than 90% of the debtors are supposed to be foreign workers, and that the data set apparently does not contain any single females.

From a statistician's point of view, using any data that does not come with a story, clear exposition of the data generation process and valid information on the meanings of the variables is at best unfortunate. Where some aspects are clearly implausible, it becomes dubious. For benchmarking prediction performance, datasets with variables that have been stripped of all meaningful information are often used. Where interest is in interpretable machine learning, such datasets become completely useless. This is one of the few data sets on credit scoring that has a meaning attached to variables and their levels. I therefore decided to try and fix the implausibility problems.

Besides the UCI Machine Learning Repository, I found that Open Data LMU (2010) also holds a copy of these data (everything in German), and the two are very closely related. The story behind these data can be found in the cited German language literature. This note provides it for the international research community, and also corrects the coding of two of the variables to become consistent with that of the other variables. The modified dataset (code for the modification see Appendix A) has been uploaded to the UCI Machine Learning Repository under the name "South German Credit" as `SouthGermanCredit.asc` (zipped with R code for reading it). This note provides a code table and background information on the data. The code table can also be used as a corrected coding for the Statlog German credit data. Appendices B and C provide code for reading both the newly contributed `SouthGermanCredit.asc` and the 1994 data file `german.data` with human readable factor levels. Appendix D lists all variables with extended verbal explanations, where variable names do not speak for themselves.

3 The “South German Credit” data

3.1 Background information

Häußler (1979, 1981) and Fahrmeir and Hamerle (1981, 1984) gave an account on the story behind the data, which is summarized here: the data are 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, which had about 500 branches, among them both urban and rural ones. Bad credits are heavily oversampled, in order to acquire sufficient information for discriminating them from good ones; the sources report that the actual prevalence of bad credits is around 5%; thus, in a Bayesian context, 5% might be used as a prior probability for a credit being bad. As suggested with the Statlog German credit data, one might consider misclassification cost, and it has been suggested to allocate the cost for misclassifying a bad risk as good to be five times as high than the cost for misclassifying a good risk as bad.

The credits are part of normal bank business, i.e. all debtors must have passed some checks of creditworthiness before being granted the credit; this, of course poses a severe limitation for the usability of the data in support of credit decisions for general requests (as was also noted e.g. by Häußler 1979). According to Fahrmeir and Hamerle (1984), customers with “good” credits perfectly complied with the conditions of the contract while customers with “bad” credits did not comply with the contract as required.

3.2 The data

The 20 explanatory variables in the data set originally contained seven quantitative and 13 categorical variables. However, four of the seven quantitative variables are only available as discretized scores and consequently are only ordinal or in one case even binary. The three remaining quantitative variables are in their original units (duration in months, amount in the German currency DM, age in years); the stated credit amounts must be considered the result of an unknown monotonic transformation, as is indicated by footnotes for Appendix 1 of Häußler (1981b) and Table 2.1 of Fahrmeir and Hamerle (1984), as well as in the title of Table 1 in Hofmann (1990).

Table 1 summarizes the categorical variables. The table has been prepared in order to support easy comparison to Table 2.1 from Fahrmeir and Hamerle (1984), and at the same time facilitate comparison to Fahrmeir and Hamerle (1981) with their variable labels M1 to M20 and to the Statlog German credit data with their attribute names A1 to A20. Factor names have been taken from R package **evtree** (Grubinger, Zeileis and Pfeiffer 2014); although these are more or less self-explanatory, Appendix D lists them together with extended explanations, where necessary. The factor levels have also largely been taken from package **evtree** and coincide with those given in the code table for the Statlog German credit data. In a few cases with particularly long level labels, these have been shortened (for example, for the highest level of factor job). For factor personal_status_sex (column `famges`), a new level had to be specified: the level with P2=2 comprises both the original categories “female : divorced/separated/married” and “male : single”; in the interest of brevity, the category has been named “female : non-single or male : single”; however, it is an open question whether female widows would belong into this category as female non-singles or into the “female : single” category, because the original code table did not mention widowed females at all. The three quantitative variables `laufzeit` = duration (A2, M3), `hoehe` = amount (A5, M2) and `alter` = age (A13, M1) are not included in the table. The response variable `kredit` = credit_risk is also not included in the table. Its levels are coded as `0` for good risks and `1` for bad risks.

Table 2.1 from Fahrmeir and Hamerle (1984) (or its second edition, respectively) is the reference for the Credit data from Open Data LMU (2010). Table 1 reproduces this table for the South German credit data, including English language factor names and levels. The percentages of both tables coincide, which creates trust that data processing was successful. The column P2 in the table arises from an expert assessment of categories’ impact on credit worthiness, as reported in Häußler (1981a). The data are provided with R code for reading them.

Table 1: Distribution of categorical predictor variables for the South German Credit data, separately for good and bad credit risks

Column	Variable name	Level	P2	bad	good
`laufkont`	status	no checking account	1	45.00	19.86
	(A1, M9)	... < 0 DM	2	35.00	23.43
		0<= ... < 200 DM	3	4.67	7.00
		... >= 200 DM / salary for at least 1 year	4	15.33	49.71
`moral`	credit_history	delay in paying off in the past	0	8.33	2.14
	(A3, M15)	critical account/other credits elsewhere	1	9.33	3.00
		no credits taken/all credits paid back duly	2	56.33	51.57
		existing credits paid back duly till now	3	9.33	8.57
		all credits at this bank paid back duly	4	16.67	34.71
`verw`	purpose	others	0	29.67	20.71
	(A4, M16)	car (new)	1	5.67	12.29
		car (used)	2	19.33	17.57
		furniture/equipment	3	20.67	31.14
		radio/television	4	1.33	1.14
		domestic appliances	5	2.67	2.00
		repairs	6	7.33	4.00
		education	7	0.00	0.00
		vacation	8	0.33	1.14
		retraining	9	11.33	9.00
		business	10	1.67	1.00
`sparkont`	savings	unknown/no savings account	1	72.33	55.14
	(A6, M7)	... < 100 DM	2	11.33	9.86
		100 <= ... < 500 DM	3	3.67	7.43
		500 <= ... < 1000 DM	4	2.00	6.00
		... >= 1000 DM	5	10.67	21.57
`beszeit`	employment_duration	unemployed	1	7.67	5.57
	(A7, M6)	< 1 yr	2	23.33	14.57
		1 <= ... < 4 yrs	3	34.67	33.57
		4 <= ... < 7 yrs	4	13.00	19.29
		>= 7 yrs	5	21.33	27.00
`rate`	installment_rate	>= 35	1	11.33	14.57
	(A8, M8)	25 <= ... < 35	2	20.67	24.14
		20 <= ... < 25	3	15.00	16.00
		< 20	4	53.00	45.29
`famges`	personal_status_sex	male : divorced/separated	1	6.67	4.29
	(A9, M17)	female : non-single or male : single	2	36.33	28.71
		male : married/widowed	3	48.67	57.43
		female : single	4	8.33	9.57
`buerge`	other_debtors	none	1	90.67	90.71
	(A10, M10)	co-applicant	2	6.00	3.29
		guarantor	3	3.33	6.00
`wohnzeit`	present_residence	< 1 yr	1	12.00	13.43
	(A11, M5)	1 <= ... < 4 yrs	2	32.33	30.14
		4 <= ... < 7 yrs	3	14.33	15.14
		>= 7 yrs	4	41.33	41.29

Table 1: Distribution of categorical predictor variables for the South German Credit data, separately for good and bad credit risks *(continued)*

Column	Variable name	Level	P2	bad	good
`verm`	property	unknown / no property	1	20.00	31.71
	(A12, M12)	car or other	2	23.67	23.00
		building soc. savings agr./life insurance	3	34.00	32.86
		real estate	4	22.33	12.43
`weitkred`	other_installment_plans	bank	1	19.00	11.71
	(A14, M13)	stores	2	6.33	4.00
		none	3	74.67	84.29
`wohn`	housing	for free	1	23.33	15.57
	(A15, M14)	rent	2	62.00	75.43
		own	3	14.67	9.00
`bishkred`	number_credits	1	1	66.67	61.86
	(A16, M4)	2-3	2	30.67	34.43
		4-5	3	2.00	3.14
		>= 6	4	0.67	0.57
`beruf`	job	unemployed/unskilled - non-resident	1	2.33	2.14
	(A17, M11)	unskilled - resident	2	18.67	20.57
		skilled employee/official	3	62.00	63.43
		manager/self-empl./highly qualif. employee	4	17.00	13.86
`pers`	people_liable	3 or more	1	15.33	15.57
	(A18, M20)	0 to 2	2	84.67	84.43
`telef`	telephone	no	1	62.33	58.43
	(A19, M19)	yes (under customer name)	2	37.67	41.57
`gastarb`	foreign_worker	yes	1	1.33	4.71
	(A20, M18)	no	2	98.67	95.29

Note:
For coding the Statlog German credit data, levels have to be switched
for columns `pers` ('people_liable', attribute A18) and `gastarb` ('foreign_worker', attribute A20).
Column P2 contains the code for the level; it is based on expert assessment, as reported in Häußler (1981a).
The Mx variable names refer to those used in Fahrmeir and Hamerle (1981) and Häußler (1979). Häußler (1981a,b) used different Mx names.

4 The two predecessor data sets and their relation to the "South German Credit" data

4.1 German credit data from Open Data LMU

Open Data LMU (2010) provided a German credit data set with purely numeric content and attributed it to several papers and a book, among them and Fahrmeir and Hamerle (1981, 1984). An excerpt from the 1984 book (Table 2.1, which is also in the 2nd edition) was provided as coding information; Fahrmeir and Hamerle (1984, 1st ed.) even listed the entire dataset in its Appendix C. The data themselves can be traced back to Häußler (1979, 1981a and b), who processed them as part of his PhD thesis.

All content and background information is in German. In fact, the coding information on columns `pers` (people_liable, number of people who financially depend on the debtor: 0-2 vs. 3 or more) and `gastarb` (foreign worker, yes or no) does not correspond to what is actually in the data, which can be verified using relative frequencies of factor levels that are printed in Table 2.1. After transforming the data values for these two variables by subtracting them from 3, the coding is as stated in the table.

The "South German Credit" data are fully compliant with Table 2.1 from Fahrmeir (1984). `SouthGermanCredit.asc` is identical with `kredit.asc` from Open Data LMU (2010), except for moving the first column to the last position and for the above-mentioned coding changes in columns `pers` and `gastarb` (see Appendix A). The original German column names have been kept as a courtesy to the data donors, and because they are nice and short.

4.2 The Statlog German credit data

The UCI Machine Learning Repository contains

- a coded data set with 7 quantitative and 13 categorical variables (`german.data`),
- a coding table for these data (`german.doc`), which will be corrected by this note,
- a further data set with post-processed categorical variables (`german.data-numeric`, not considered in this note).

The data have been contributed as part of a dataset collection created by the Statlog EU project (https://cordis.europa.eu/project/rcn/8791/factsheet/en); Prof. Dr. Hans-Joachim Hofmann, a retired professor from Hamburg university is listed as the data donor. Information on data generation or context is missing, as was also recently criticized by an anonymous blogger on reddit (Anonymous 2019). A derived version of the data is also available from Kaggle at https://www.kaggle.com/uciml/german-credit; the Kaggle contributor criticized the original data: "It is almost impossible to understand the original dataset due to its complicated system of categories and symbols. Thus, I wrote a small Python script to convert it into a readable CSV file. Several columns are simply ignored, because in my opinion either they are not important or their descriptions are obscure." Unfortunately, the Kaggle German credit data are also affected by the wrong code table of the Statlog German credit data; even the sex column is incorrect, because the variable "personal status and sex" contains a combined category that includes both single males and non-single females, which is not properly reflected by the code table provided with the Statlog German credit data.

The complaint about the complicated system is presumably due to inclusion of an attribute identifier in the data table itself: for all 13 intrinsically categorical variables, the numeric category code is preceded by the attribute identifier, e.g. the categories 0 to 4 for attribute A3 would be denoted as A30 to A34. They are thus character strings, and care has to be taken to process them adequately, even if correct codings are used. The codes in the raw data are largely identical between the German credit data from Open Data LMU (2010) and the Statlog German credit data (see below for more detailed considerations). Thus, for all attributes except A18 and A20, Table 1 can be used as a code book for the Statlog German credit data; for A18 and A20, the levels have to be swapped.

Anonymous (2019) identified a paper by Hofmann (1990), which is hard to come by from universities outside of German speaking countries (and inconvenient even from within Germany). That paper attributes the data to papers by Häußler (1979, 1981a), and to the printed data appendix of Fahrmeir and Hamerle (1984); Hofmann has typed in the first 1000 data rows from the appendix of that book (he stated that the appendix contained more than 1000 data rows and that he ignored rows 1001 and following, because the book reports on exactly 1000 cases; indeed, the 1001st printed row and the next few rows contain weird data entries; thus, ignoring everything from row 1001 onwards appears wise).

It appears that Hofmann himself worked with correctly coded data in his 1990 paper: he mentioned that he checked his typed data against Table 2.1 from Fahrmeir and Hamerle (1984), and his CART results are also plausible. Presumably, Hofmann donated the data to the Statlog EU project, and a representative of Strathclyde university took care of the donation to the UCI Machine Learning Repository. The wrong code table must have been added somewhere along the way. Mixups are dramatic: attributes A1, A3, A4, A6, A9, A12, A15 and A20 have their levels mixed up, and for attribute A9, the two collapsed levels (female : divorced/separated/married with male : single) are not properly reflected in the coding, which leads to the wrong assumption that a sex column can be derived from that attribute.

Hofmann also warned that there might be some typos in his data; a check by merging the German Credit data from Open Data LMU (2010) with the Statlog German credit data reveals that there are surprisingly few discrepancies: the two data sets have been merged based on the three truly quantitative variables; only 10 records could not be matched because of typos either in credit amount (four cases, small changes only) or age (six cases, at most a difference of 10 years). Among the 990 matched records, there are 34 mismatches for attribute A14, one mismatch for A15 and 10 mismatches for A20. Thus, the Statlog German credit data, when used with a corrected code table, are almost identical to the German credit data from Open Data LMU (2010) and thus also to the "South German Credit" data. As was mentioned before, the latter differ in that they use the P2 scores (according to Häußler 1981a) for coding *all* categorical variables, including for `pers`=A18=`persons_liable` and `gastarb`=A20=`foreign_worker`.

5 Summary and final comments

The "South German Credit" data are meaningful, however very old (1973 to 1975), credit scoring data from southern Germany. There are 1000 rows with 20 predictor variables (quantitative, ordinal and nominal variables) and one binary response variable. The data can be used for regression or classification applications. Even nominal predictor variables have been coded with scores (P2) that can be considered ordinal, because they are based on an expert assessment of which categories lead to better (=larger number) or worse (=smaller number) credit risk. The coding table for these data (see Table 1) can also be used as a corrected coding table for the UCI Machine Learning Repository's German credit data (contributed by Hofmann in 1994).

Locating the fixes for the code table of the Statlog German credit data required some persistence, since the data did not come with any documentation beyond the stated erroneous code table. Anonymous (2019) helped by locating the Hofmann (1990) paper, which in turn helped me find the right search terms for locating the Open Data LMU (2010) version of the German credit data (even before I managed to obtain a copy of the Hofmann (1990) paper). These difficulties underline that it is very important for published research data to be well-documented.

I would like to close this note with an appeal to data donors to find ways to donate data with meaningful content –stripping a dataset from all human-interpretable information by providing purely technical labels certainly helps with satisfying confidentiality needs. While the resulting data may still be usable for benchmarking algorithm performance, the increasingly important field of research on interpretable machine learning needs data with meaningful content.

References

Anonymous (2019, reddit user name SoFarFromHome; accessed November 27, 2019). https://www.reddit.com/r/MachineLearning/comments/d7zsqf/d_the_german_credit_rating_data_set_widely_used/.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fahrmeir, L. and Hamerle, A. (1981, in German). Kategoriale Regression in der betrieblichen Planung. *Zeitschrift für Operations Research* **25**, B63-B78.

Fahrmeir, L. and Hamerle, A. (1984, in German). *Multivariate Statistische Verfahren* (1st ed., Ch.8 and Appendix C). De Gruyter, Berlin.

Grubinger, T. Zeileis, A. and Pfeiffer, K.-P. (2014). **evtree**: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. *Journal of Statistical Software* **61**(1), 1-29. http://www.jstatsoft.org/v61/i01/.

Häußler, W.M. (1979, in German). Empirische Ergebnisse zu Diskriminationsverfahren bei Kreditscoringsystemen. *Zeitschrift für Operations Research* **23**, B191-B210.

Häußler, W.M. (1981a, in German). Methoden der Punktebewertung bei Kreditscoringsystemen. *Zeitschrift für Operations Research*, Series B, **25**, B79-B94.

Häußler, W.M. (1981b, in German). Über Verfahren der Punktebewertung und Diskrimination mit Anwendung auf Kreditscoringsysteme. PhD thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig. (Published by Fritz Knapp Verlag, Frankfurt am Main, as *Punktebewertungen bei Kreditscoringsystemen.*)

Hofmann, H.J. (1990, in German). Die Anwendung des CART-Verfahrens zur statistischen Bonitätsanalyse von Konsumentenkrediten. *Zeitschrift für Betriebswirtschaft* **60**, 941–962.

Open data LMU (2010; accessed Nov 27 2019; in German). Kreditscoring zur Klassifikation von Kreditnehmern. https://doi.org/10.5282/ubm/data.23.

6 Appendix A: R Code for creating SouthGermanCredit.asc from kredit.asc

```r
LMU <- read.table("https://data.ub.uni-muenchen.de/23/2/kredit.asc", header=TRUE)
### recode pers and gastarb to the stated P2 coding
LMU$pers <- 3 - LMU$pers
LMU$gastarb <- 3 - LMU$gastarb
### reorder columns so that credit_risk is last
LMU <- cbind(LMU[,-1], kredit=LMU$kredit)
write.table(LMU, file="SouthGermanCredit.asc",
            row.names = FALSE, quote=FALSE)
```

7 Appendix B: R Code for reading SouthGermanCredit.asc

```r
## make sure to use your own path
dat <- read.table("GermanCredit/SouthGermanCredit.asc", header=TRUE)
## everything is numeric

## dat contains numbers for all variables.

## variable names from Fahrmeir/Hamerle book
nam_fahrmeirbook <- c("laufkont", "laufzeit", "moral", "verw",
                      "hoehe", "sparkont", "beszeit", "rate",
                      "famges", "buerge", "wohnzeit", "verm",
                      "alter", "weitkred", "wohn", "bishkred",
                      "beruf", "pers", "telef", "gastarb",
                      "kredit")
nam_evtree <- c("status", "duration", "credit_history",
      "purpose", "amount", "savings", "employment_duration",
      "installment_rate", "personal_status_sex",
      "other_debtors", "present_residence", "property",
      "age", "other_installment_plans", "housing",
      "number_credits", "job", "people_liable", "telephone",
      "foreign_worker", "credit_risk")
names(dat) <- nam_evtree

## make factors for all except the numeric variables
## make sure that even empty level of factor purpose = verw (dat[[4]]) is included
for (i in setdiff(1:21, c(2,4,5,13)))
  dat[[i]] <- factor(dat[[i]])
## factor purpose
dat[[4]] <- factor(dat[[4]], levels=as.character(0:10))

## assign level codes
## make intrinsically ordered factors into class ordered
levels(dat$credit_risk) <- c("bad", "good")
levels(dat$status) = c("no checking account",
                   "... < 0 DM",
                   "0<= ... < 200 DM",
                   "... >= 200 DM / salary for at least 1 year")
## "critical account/other credits elsewhere" was
## "critical account/other credits existing (not at this bank)",
levels(dat$credit_history) <- c(
  "delay in paying off in the past",
  "critical account/other credits elsewhere",
  "no credits taken/all credits paid back duly",
  "existing credits paid back duly till now",
  "all credits at this bank paid back duly")
levels(dat$purpose) <- c(
  "others",
  "car (new)",
  "car (used)",
  "furniture/equipment",
  "radio/television",
  "domestic appliances",
  "repairs",
  "education",
  "vacation",
  "retraining",
  "business")
levels(dat$savings) <- c("unknown/no savings account",
                         "... <  100 DM",
                         "100 <= ... <  500 DM",
                         "500 <= ... < 1000 DM",
                         "... >= 1000 DM")
levels(dat$employment_duration) <-
                  c(  "unemployed",
                      "< 1 yr",
                      "1 <= ... < 4 yrs",
                      "4 <= ... < 7 yrs",
                      ">= 7 yrs")
dat$installment_rate <- ordered(dat$installment_rate)
levels(dat$installment_rate) <- c(">= 35",
                                  "25 <= ... < 35",
                                  "20 <= ... < 25",
                                  "< 20")
levels(dat$other_debtors) <- c(
  "none",
  "co-applicant",
  "guarantor"
)
## female : nonsingle was female : divorced/separated/married
levels(dat$personal_status_sex) <- c(
  "male : divorced/separated",
  "female : non-single or male : single",
  "male : married/widowed",
  "female : single")
dat$present_residence <- ordered(dat$present_residence)
levels(dat$present_residence) <- c("< 1 yr",
                                   "1 <= ... < 4 yrs",
                                   "4 <= ... < 7 yrs",
                                   ">= 7 yrs")
levels(dat$property) <- c(
  "unknown / no property",
  "car or other",
  "building soc. savings agr. / life insurance",
  "real estate"
)
levels(dat$other_installment_plans) <- c(
  "bank",
  "stores",
  "none"
)
levels(dat$housing) <- c("for free", "rent", "own")
dat$number_credits <- ordered(dat$number_credits)
levels(dat$number_credits) <- c("1", "2-3", "4-5", ">= 6")
## manager/self-empl/highly qualif. employee  was
##   management/self-employed/highly qualified employee/officer
levels(dat$job) <- c(
  "unemployed/unskilled - non-resident",
  "unskilled - resident",
  "skilled employee/official",
  "manager/self-empl/highly qualif. employee"
)
levels(dat$people_liable) <- c("3 or more", "0 to 2")
levels(dat$telephone) <- c("no", "yes (under customer name)")
levels(dat$foreign_worker) <- c("yes", "no")
```

8 Appendix C: R Code for reading `german.data`

```
## make sure to use your own path, or the appropriate URL
dat <- read.table("GermanCredit/german.data", stringsAsFactors = FALSE)
## remove Ax portions from string columns
for (i in 1:ncol(dat))
  if (is.character(dat[[i]]))
    dat[[i]] <- as.integer(gsub(paste0("A",i), "", dat[[i]]))

## Now dat contains numbers for all variables.
## The rest is as in Appendix A,
##     except for  credit_risk,
##                 attribute 18: people_liable
##             and attribute 20: foreign_worker

nam_evtree <- c("status", "duration", "credit_history",
      "purpose", "amount", "savings", "employment_duration",
      "installment_rate", "personal_status_sex",
      "other_debtors", "present_residence", "property",
      "age", "other_installment_plans", "housing",
      "number_credits", "job", "people_liable", "telephone",
      "foreign_worker", "credit_risk")
names(dat) <- nam_evtree

## make factors for all except the numeric variables
## make sure that even empty level of factor purpose = verw (dat[[4]]) is included
for (i in setdiff(1:21, c(2,4,5,13)))
  dat[[i]] <- factor(dat[[i]])
## factor purpose
dat[[4]] <- factor(dat[[4]], levels=as.character(0:10))

## assign level codes
## make intrinsically ordered factors into class ordered
levels(dat$credit_risk) <- c("good", "bad")

###### use code from Appendix A for status to job
###### last three again from here:
```

```r
levels(dat$people_liable) <- c("0 to 2", "3 or more")
levels(dat$telephone) <- c("no", "yes (under customer name)")
levels(dat$foreign_worker) <- c("no", "yes")
```

Appendix D: The 21 variables of the (South) German credit data

Column name: `laufkont`
Variable name: status
Content: status of the debtor's checking account with the bank (categorical)

Column name: `laufzeit`
Variable name: duration
Content: credit duration in months (quantitative)

Column name: `moral`
Variable name: credit_history
Content: history of compliance with previous or concurrent credit contracts (categorical)

Column name: `verw`
Variable name: purpose
Content: purpose for which the credit is needed (categorical)

Column name: `hoehe`
Variable name: amount
Content: credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)

Column name: `sparkont`
Variable name: savings
Content: debtor's savings (categorical)

Column name: `beszeit`
Variable name: employment_duration
Content: duration of debtor's employment with current employer (ordinal; discretized quantitative)

Column name: `rate`
Variable name: installment_rate
Content: credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)

Column name: `famges`
Variable name: personal_status_sex
Content: combined information on sex and marital status; categorical; sex cannot be recovered from the variable, because male singles and female non-singles are coded with the same code (2); female widows cannot be easily classified, because the code table does not list them in any of the female categories

Column name: `buerge`
Variable name: other_debtors
Content: Is there another debtor or a guarantor for the credit? (categorical)

Column name: `wohnzeit`
Variable name: present_residence
Content: length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)

Column name: `verm`
Variable name: property
Content: the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if codes 3 or 4 are not applicable and there is a car or any other relevant property that does not fall under variable sparkont. (ordinal)

Column name: `alter`
Variable name: age
Content: age in years (quantitative)

Column name: `weitkred`
Variable name: other_installment_plans
Content: installment plans from providers other than the credit-giving bank (categorical)

Column name: `wohn`
Variable name: housing
Content: type of housing the debtor lives in (categorical)

Column name: `bishkred`
Variable name: number_credits
Content: number of credits including the current one the debtor has (or had) at this bank (ordinal, discretized quantitative); contrary to Fahrmeir and Hamerle's (1984) statement, the original data values are not available.

Column name: `beruf`
Variable name: job
Content: quality of debtor's job (ordinal)

Column name: `pers`
Variable name: people_liable
Content: number of persons who financially depend on the debtor (i.e., are entitled to maintenance) (binary, discretized quantitative)

Column name: `telef`
Variable name: telephone
Content: Is there a telephone landline registered on the debtor's name? (binary; remember that the data are from the 1970s)

Column name: `gastarb`
Variable name: foreign_worker
Content: Is the debtor a foreign worker? (binary)

Column name: `kredit`
Variable name: credit_risk
Content: Has the credit contract been complied with (good) or not (bad) ? (binary)