21.2 GERMAN CREDIT

GermanCredit.csv is the dataset for this case study.

Background

Money-lending has been around since the advent of money; it is perhaps the world's second-oldest profession. The systematic evaluation of credit risk, though, is a relatively recent arrival, and lending was largely based on reputation and very incomplete data. Thomas Jefferson, the third President of the United States, was in debt throughout his life and unreliable in his debt payments, yet people continued to lend him money. It wasn't until the beginning of the 20th century that the Retail Credit Company was founded to share information about credit. That company is now Equifax, one of the big three credit scoring agencies (the other two are Transunion and Experion).

Individual and local human judgment are now largely irrelevant to the credit reporting process. Credit agencies and other big financial institutions extending credit at the retail level collect huge amounts of data to predict whether defaults or other adverse events will occur, based on numerous customer and transaction information.

Data

This case deals with an early stage of the historical transition to predictive modeling, in which humans were employed to label records as either good or poor credit. The German Credit dataset[2] has 30 variables and 1000 records, each record being a prior applicant for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases). Table 21.2 shows the values of these variables for the first four records. All the variables are explained in Table 21.3. New applicants for credit can also be evaluated on these 30 predictor variables and classified as a good or a bad credit risk based on the predictor values.

The consequences of misclassification have been assessed as follows: The costs of a false positive (incorrectly saying that an applicant is a good credit risk) outweigh the benefits of a true positive (correctly saying that an applicant is a good credit risk) by a factor of 5. This is summarized in Table 21.4. The opportunity cost table was derived from the average net profit per loan as shown in Table 21.5. Because decision makers are used to thinking of their decision in terms of net profits, we use these tables in assessing the performance of the various models.

[2]This dataset is available from ftp.ics.uci.edu/pub/machine-learning-databases/statlog.

TABLE 21.2 FIRST FOUR RECORDS FROM GERMAN CREDIT DATASET

OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT	INSTALL_RATE	MALE_DIV	MALE_SINGLE	MALE_MAR_WID	CO-APPLICANT	GUARANTOR
1	0	6	4	0	0	0	1	0	0	1169	4	4	4	0	1	0	0	0
2	1	48	2	0	0	0	1	0	0	5951	0	2	2	0	0	0	0	0
3	3	12	4	0	0	0	0	1	0	2096	0	3	2	0	1	0	0	0
4	0	42	2	0	0	1	0	0	0	7882	0	3	2	0	1	0	0	1

PRESENT_RESIDENT	REAL_ESTATE	PROP_UNKN_NONE	AGE	OTHER_INSTALL	RENT	OWN_RES	NUM_CREDITS	JOB	NUM_DEPENDENTS	TELEPHONE	FOREIGN	RESPONSE
4	1	0	67	0	0	1	2	2	1	1	0	1
2	1	0	22	0	0	1	1	2	1	0	0	0
3	1	0	49	0	0	1	1	1	2	0	0	1
4	0	0	45	0	0	0	1	2	2	0	0	1

(Data adapted from German Credit)

TABLE 21.3 VARIABLES FOR THE GERMAN CREDIT DATASET

Variable number	Variable name	Description	Variable type	Code description
1	OBS#	Observation numbers	Categorical	Sequence number in dataset
2	CHK_ACCT	Checking account status	Categorical	0: <0 DM 1: 0–200 DM 2 : >200 DM 3: No checking account
3	DURATION	Duration of credit in months	Numerical	
4	HISTORY	Credit history	Categorical	0: No credits taken 1: All credits at this bank paid back duly 2: Existing credits paid back duly until now 3: Delay in paying off in the past 4: Critical account
5	NEW_CAR	Purpose of credit	Binary	Car (new), 0: no, 1: yes
6	USED_CAR	Purpose of credit	Binary	Car (used), 0: no, 1: yes
7	FURNITURE	Purpose of credit	Binary	Furniture/equipment, 0: no, 1: yes
8	RADIO/TV	Purpose of credit	Binary	Radio/television, 0: no, 1: yes

(continued)

TABLE 21.3 *(CONTINUED)*

Variable number	Variable name	Description	Variable type	Code description
9	EDUCATION	Purpose of credit	Binary	Education, 0: no, 1: yes
10	RETRAINING	Purpose of credit	Binary	Retraining, 0: no, 1: yes
11	AMOUNT	Credit amount	Numerical	
12	SAV_ACCT	Average balance in savings account	Categorical	0: $<$100 DM 1 : 101—500 DM 2 : 501—1000 DM 3 : $>$1000 DM 4 : Unknown/no savings account
13	EMPLOYMENT	Present employment since	Categorical	0 : Unemployed 1: $<$1 year 2: 1—3 years 3: 4—6 years 4: $\geq$7 years
14	INSTALL_RATE	Installment rate as % of disposable income	Numerical	
15	MALE_DIV	Applicant is male and divorced	Binary	0: no, 1: yes
16	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
17	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18	CO-APPLICANT	Application has a coapplicant	Binary	0: No, 1: Yes
19	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20	PRESENT_RESIDENT	Present resident since (years)	Categorical	0: $\leq$1 year 1: 1—2 years 2: 2—3 years 3: $\geq$3 years
21	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
22	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23	AGE	Age in years	Numerical	
24	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
25	RENT	Applicant rents	Binary	0: No, 1: Yes
26	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes

(continued)

TABLE 21.3 *(CONTINUED)*

Variable number	Variable name	Description	Variable type	Code description
27	NUM_CREDITS	Number of existing credits at this bank	Numerical	
28	JOB	Nature of job	Categorical	0 : Unemployed/ unskilled— non-resident 1 : Unskilled— resident 2 : Skilled employee/ official 3 : Management/ self-employed/ highly qualified employee/officer
29	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

The original dataset had a number of categorical variables, some of which were transformed into a series of binary variables and some ordered categorical variables were left as is, to be treated as numerical. (Data adapted from German Credit)

TABLE 21.4 OPPORTUNITY COST TABLE (DEUTSCHE MARKS)

	Actual	
Predicted (decision)	**Good**	**Bad**
Good (accept)	0	500
Bad (reject)	100	0

(Data adapted from Deutsche Marks)

TABLE 21.5 AVERAGE NET PROFIT (DEUTSCHE MARKS)

	Actual	
Predicted (decision)	**Good**	**Bad**
Good (accept)	100	−500
Bad (reject)	0	0

(Data adapted from Deutsche Marks)

Assignment

1. Review the predictor variables and guess what their role in a credit decision might be. Are there any surprises in the data?
2. Divide the data into training and validation partitions, and develop classification models using the following data mining techniques: logistic regression, classification trees, and neural networks.
3. Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the highest net profit?
4. Let us try and improve our performance. Rather than accept the default classification of all applicants' credit status, use the estimated probabilities (propensities) from the logistic regression (where *success* means 1) as a basis for selecting the best credit risks first, followed by poorer-risk applicants. Create a vector containing the net profit for each record in the validation set. Use this vector to create a decile-wise lift chart for the validation set that incorporates the net profit.
   a. How far into the validation data should you go to get maximum net profit? (Often, this is specified as a percentile or rounded to deciles.)
   b. If this logistic regression model is used to score to future applicants, what "probability of success" cutoff should be used in extending credit?