Case: German Credit

The German Credit data set (available at ftp.ics.uci.edu/pub/machine-learning-databases/statlog/) contains observations on 30 variables for 1000 past applicants for credit. Each applicant was rated as “good credit” (700 cases) or “bad credit” (300 cases).

New applicants for credit can also be evaluated on these 30 "predictor" variables. We want to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. All the variables are explained in Table 1.1. (Note: The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be appropriately handled by XLMiner. Several ordered categorical variables have been left as is; to be treated by XLMiner as numerical. The data has been organized in the spreadsheet German CreditI.xls)

Var. #	Variable Name	Description	Variable Type	Code Description
1.	OBS#	Observation No.	Categorical	Sequence Number in data set
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 till 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
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 : 100<= ... < 500 DM 2 : 500<= ... < 1000 DM 3 : =>1000 DM 4 : unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1: < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years

				4 : >= 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 co-applicant	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: <= 1 year
				1<…<=2 years
				2<…<=3 years
				3:>4years
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
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

Table 1.1 Variables for the German Credit data.

Table 1.2, below, shows the values of these variables for the first several records in the case.

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

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

Table 1.2 The data (first several rows)

The consequences of misclassification have been assessed as follows: the costs of a false positive (incorrectly saying an applicant is a good credit risk) outweigh the cost of a false negative (incorrectly saying an applicant is a bad credit risk) by a factor of five. This can be summarized in the following table.

	Predicted (Decision)		
Actual		Good (Accept)	Bad (Reject)
	Good	0	100 DM
	Bad	500 DM	0

Table 1.3 Opportunity Cost Table (in deutch Marks)

The opportunity cost table was derived from the average net profit per loan as shown below:

	Predicted (Decision)		
Actual		Good (Accept)	Bad (Reject)
	Good	100 DM	0
	Bad	- 500 DM	0

Table 1.4 Average Net Profit

Let us use this table in assessing the performance of the various models because it is simpler to explain to decision-makers who are used to thinking of their decision in terms of net profits.

Assignment

1. Review the predictor variables and guess from their definition at what their role might be in a credit decision. Are there any surprises in the data?

2. Divide the data randomly into training (60%) and validation (40%) partitions, and develop classification models using the following data mining techniques in XLMiner:
   - Logistic regression
   - Classification trees
   - Neural networks
   - Discriminant Analysis.

3. Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. For the logistic regression model use a cutoff “predicted probability of success” ("success"=1) of 0.5. Which technique gives the most net profit on the validation data?

4. Let's see if we can improve our performance by changing the cutoff. Rather than accepting XLMiner's initial classification of everyone's credit status, let's use the "predicted probability of success" in logistic regression as a basis for selecting the best credit risks first, followed by poorer risk applicants.

a. Sort the validation data on "predicted probability of success."
b. For each validation case, calculate the actual cost/gain of extending credit.
c. Add another column for cumulative net profit.
d. How far into the validation data do you go to get maximum net profit? (Often this is specified as a percentile or rounded to deciles.)
e. If this logistic regression model is scored to future applicants, what "probability of success" cutoff should be used in extending credit?