CREDIT RISK PREDICTION

Background

When a bank receives a loan application, based on the applicant's profile bank has to decide whether to go ahead with the loan approval or not. Two types of risks are associated with the bank's decision –

> If the applicant is a good credit risk , i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank.
>
> If the applicant is a bad credit risk, i.e is not likely to repay the loan , then approving the loan to the person results in a financial loss to the bank.

Business Objective

To minimize loss from the bank's perspective , the bank needs a decision rule regarding who to give approval & who not to. An applicant's demographic and social-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application

This assignment is related to building a logistic regression model on credit data. It contains data on a 1000 customers of a bank, and their credit rating (Good/Bad) based on previous history. The variable response in the dataset corresponds to the risk label, 1 has been classified as bad and 2 has been classified as good.

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.

Data dictionary.

There is a total on 29 attributes are there in the dataset. Their descriptions and details have been tabulated below:

Codelist Var. #	Variable Name	Description	Variable Type	Code Description	Sheet Name
1	OBS#	Observation No.	Categorical		Part1
2	DURATION	Duration of credit in months	Numerical		Part1
3	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes	Part1
4	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes	Part1
5	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes	Part1
6	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes	Part1
7	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes	Part1
8	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes	Part1
9	INSTALL_RATE	Installment rate as % of disposable income	Numerical		Part1
10	CO-APPLICANT	Application has a co-applicant	Binary	0: No, 1: Yes	Part1
11	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes	Part1
12	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes	Part1
13	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes	Part1
14	AGE	Age in years	Numerical		Part1
15	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes	Part1
16	RENT	Applicant rents	Binary	0: No, 1: Yes	Part1
17	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes	Part1
18	NUM_CREDITS	Number of existing credits at this bank	Numerical		Part1
19	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical		Part1
20	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes	Part1
21	FOREIGN	Foreign worker	Binary	0: No, 1: Yes	Part1
22	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes	Part1
23	AMOUNT	Credit amount	Numerical		part1
24	CHK_ACCT	Checking account status	Categorical		Part2
				0 : < 0 GB 1: 0 < .. < 200 GB 2 : => 200 GB 3 : no account	
25	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	part2
26	SAV_ACCT	Average balance in savings account	Categorical	0 : < 100 GB 1 : 100<= ... < 500 GB 2 : 500<= ... < 1000 GB 3 : =>1000 GB 4 : unknown/ no savings account	part2
27	EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1: < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years 4 : >= 7 year	part 2
28	PRESENT_RESIDENT	Present resident since - years	Categorical	0: <= 1 year 1: 1<…<=2 years 2: 2<…<=3 years 3:>4years	part2
29	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	part2

Tasks to be carried out

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 machine Learning techniques in Python & R:

- Logistic regression
- Classification trees
- Neural networks

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.