DSME6756BA Homework 4

Due date : 11:59pm on Monday, 6 May

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. The data has been organized in the spreadsheet GermanCredit.xlsx. All the variables are explained in 'Codelist' worksheet of the data file. (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 Python. Several ordered categorical variables have been left as is; to be treated by Python as numerical).

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 1.

Table 1 Opportunity Cost

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

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

Table 2 Average Net Profit

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

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

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 a classification model using the logistic regression technique in Python and evaluate the model by using the confusion matrix and the ROC curve.

3. Based on the confusion matrix and the payoff matrix, what is the net profit on the validation data?

4. Let's see if we can improve our performance by changing the cutoff. Rather than accepting the above classification of everyone's credit status, let's use the "predicted probability of finding a good applicant" 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 finding a good applicant."
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?

Submission Guidelines

Please submit your work via Blackboard by the deadline. When you submit via Blackboard, you should attach your Jupyter notebook.