ST404: Applied Statistical Modelling

Assignment 3: Credit Card Data

3.0 Assignment Weighting

Assignment 3 counts for 35% of the module mark.

3.1 Deadline:

12:00 Tuesday 4th May 2021

to be submitted electronically via Moodle.

3.2 Report

For this assignment you are asked to produce a report of no more than 12 pages (excluding both the bibliography and the appendix).
The front cover of your report should give your student ID, but not your name (to allow anonymous marking).
Your report should contain the following sections:

- A short introduction section describing the problem.
- A technical methods section that can be understood by your fellow ST404 students, describing your approach to model fitting as well as the results of any diagnostics you may use.
- A results section that can be understood by the client (see below), assuming that they have no specialist knowledge of statistics, summarizing your findings.
- A bibliography.
- An appendix containing your annotated R code (you may reduce this if you have repetitive code to just one example e.g. if you plot each continuous variable you only need to include the code for one such plot).

The report should be typeset in a professional manner with appropriate margins (at least one inch), font size 11 or higher and 1.5 spacing. All figures and tables should be numbered and have captions. Pages should be numbered.

Keep in mind the advice on academic writing and the rules about referencing, plagiarism and proof-reading. Make sure that all sources used, whether online or paper-based are appropriately referenced. The assignment will be submitted to TurnItIn and any cases of potential plagiarism forwarded to the departmental academic conduct panel.

This is an individual assignment. Collaboration between students is not permitted (other than questions/answers posted on the discussion forum) and will be treated as cheating.

3.3 PROBLEM OUTLINE

You are acting for a consultancy firm and have been asked by a Taiwanese Credit Card Company[‡] to help them to predict customers who are likely to default on their credit card.

You have been provided with two sample data sets of customers who have a credit limit that is equal or greater than TN$250,000. (TN$ are Taiwanese dollars.) As this information is confidential you have been provided with a historic data set in order to demonstrate "proof of concept". If your firm is successful it will be commissioned in the future to provide modelling services for current data. The aim is to build a model that is able to predict which customers are likely to default on paying their card bills. This can be later used to build a suitable score card for customers applying for a card or for further credit on their card. It is important for the company to be able to explain the justification for refusing someone credit and therefore as well as a model that is able to predict, they are interested in the interpretation of the model.

3.3.1 Data Availability

The data are provided on Moodle as two R data sets and consist of:

Training Data: **CardT.rds**: 5000 observations

Validation Data: **CardV.rds:** 2067 observations

Both data sets contain the same variables which are listed in Table 3.1 on page 3 below. To read these in please use the `readRDS` command. If you copy the files to your R working directory, then to read in these data the commands would be:

```
CardT <- readRDS("CardT.rds")
CardV <- readRDS("CardV.rds")
```

[‡] Data extracted from "default of credit card clients Data Set" donated by I-Cheng Yeh and from 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.

Variable	Type	Details: description / Factor Level = label (meaning)		
LIMIT_BAL	Continuous	Credit Limit (NT$): it includes both the individual consumer credit and his/her family (supplementary) credit		
SEX	Factor	Gender		1=male 2=female
EDUCATION	Factor	Education Level		1 = GradSch(graduate school) 2 = Uni (university) 3 = HighSch (high school) 4 = Other (others)
MARRIAGE	Factor	Marital status		1 = married 2 = single 3 = others
AGE	Continuous	Age in years		
PAY_1	Factor	The repayment status in	September 2005	-2 = no consumption -1 = pay duly 0 = the use of revolving credit 1 = payment delay for one month 2 = payment delay for two months ⋮ 8 = payment delay for eight months 9 = payment delay for nine months and above[§]
PAY_2	Factor		August 2005	
PAY_3	Factor		July 2005	
PAY_4	Factor		June 2005	
PAY_5	Factor		May 2005	
PAY_6	Factor		April 2005	
BILL_AMT1	Continuous	Amount of bill statement in	September 2005	Taiwanese Dollars (NT$)
BILL_AMT2	Continuous		August 2005	
BILL_AMT3	Continuous		July 2005	
BILL_AMT4	Continuous		June 2005	
BILL_AMT5	Continuous		May 2005	
BILL_AMT6	Continuous		April 2005	
PAY_AMT1	Continuous	Amount paid in	September 2005	
PAY_AMT2	Continuous		August 2005	
PAY_AMT3	Continuous		July 2005	
PAY_AMT4	Continuous		June 2005	
PAY_AMT5	Continuous		May 2005	
PAY_AMT6	Continuous		April 2005	
default	binary	Default payment (dependent variable)		0 = No 1 = Yes

Table 3.1: Details of Variables in provided data sets

§ http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/ClassificationProcessCreditCardDefault.html

3.3.2 Data Analyses Required

1) You should begin your analysis with a suitable EDA of the data, this should look at univariate considerations and at the relationships between variables especially to the dependent variable. This may result in you requiring to then manipulate your data in order to fit a suitable model.
2) Using the training data, you should then build a suitable logistic regression model in order to predict which customers are likely to ***default***.
   a) You should take into consideration any information you found from your initial EDA when building your model. e.g., are there any steps that need to be taken before model fitting?
   b) You should aim to use at least two selection methods.
3) You should validate your model using suitable evaluation tools (e.g., diagnostic plots.)
4) As the dependent variable is a binary variable you should evaluate your model performance. Choose a suitable cut off point for the predicted probability of default to make a binary prediction "likely to default" or "unlikely to default" then calculate the false positive and false negative rates using the validation data. You may also produce a ROC chart.
5) You should discuss the interpretation, success and limitations of your final model.

3.4 ASSESSMENT CRITERIA

Content (30%)
Under this heading comes an assessment of:

- your background reading, and your appreciation of the context;
- the validity, accuracy and relevance of the results of statistical analysis;
- the conceptual difficulty of the material you have discussed;

Understanding (30%)
Understanding shows in the way in which you write. Your understanding of the topic is reflected in:

- the clarity of the explanation you offer to your reader;
- the degree of accuracy with which you use statistical notation and terminology;
- the construction of logically sound arguments.

Originality (20%)

- Originality of presentation: you are all analysing the same data, but as you read round the topic and develop your understanding of it you hopefully will find yourself having your own version of the story to tell.
- Originality of content: a practical project will inevitably contain new material, for example: the results of the data analysis, computer programs written and similar individual aspects.

Presentation (20%)

- Is your language appropriate for the intended target audience?
- Does your choice of structure for the narrative help the reader gain understanding?
- Are your sentences clear and coherent without distracting grammatical and spelling errors?
- Are tables, graphs and diagrams relevant, easy to follow and well positioned?
- Are sources appropriately referenced?

Penalties:

- Late submission (-5% per working day)
- Over page limit (-5% per page)
- Not using appropriate layout (-5%)

3.4.1 Additional Points

Within the contents and understanding sections above, various aspects of your analyses as outlined in 1) through to 5) of section 3.3.2 on page 4 should be covered. There is no specific weighting to the 5 steps listed and the hope is that you will demonstrate good choices in what to include and discuss. However, you should aim to present a good balance of these 5 stages, i.e. in the main your report should cover all of them. It may be that your report will focus slightly more on some of these aspects than others, but one of the criteria that is required for a good grade from this assignment is a suitable balance. You therefore should avoid focusing on merely one or two of these and you will need to select a suitable mix of your findings, results and discussion. The aim is to guide the reader through the steps you have used and your findings.

3.5 ADDITIONAL RESOURCES

Some additional code is provided that will enable you to produce some empirical logit plots in order to understand the relationship between the predictors and the dependent variable ***default***. The code illustrates how this may be used for the Pima Indians data. In addition, the code will illustrate how a ROC chart may be produced for the Pima Indians data. It is not required that you necessarily make use of this, nor should your analysis be limited to this code. It is provided as an extra resource to illustrate some of the specialised aspects you may wish to consider.