MATH60131/70131 CONSUMER CREDIT RISK MODELLING - PROJECT 2022

In this exercise you will be assigned a data set of personal loans and you will build 2 scorecards using the techniques taught on the Consumer Credit Risk Modelling course. You will use the R statistical language for your work and report the methods you use and your results in a written report.
Total contribution to assessment: 25%.

DEADLINE

Your coursework must be submitted by 1pm on Monday 5 December 2022.

INSTRUCTIONS

1. On Blackboard Learn, in the *Coursework* folder, you will see ten data files of the form `LCdata_`*n*`.RData`. For this coursework, download and use the file where *n* = the last digit of your College ID. Load this file into your R using the `load` function; eg

   ```
   load("H:\\data\\LCdata_1.Rdata")
   ```

   will load data from folder `data` into a data frame named `D1`. The data is described in Appendix A below.

2. First, you will build a logistic regression model of non-default, where default is given by the *def_flag* variable and with *only* the predictor variables: loan_amnt, grade, emp_length_p, term and addr_state. Analyse each of these variables and decide which transformation or processing is required to include them in your model, if any.

3. Split the data randomly into a training data set and a test data set, with the ratio 2:1 of observations in each.

4. Build a scorecard using a single logistic regression model with the outcome and predictor variables as described and processed in step 2.
   - Do not include interaction terms.
   - Build the model using only the training data set.

5. Interpret your scorecard, considering which predictor variables are important in your model and how they are associated with creditworthiness, in terms of direction and statistical significance of association. Use a significance level of 0.01.

6. Use your scorecard to construct the ROC curve and compute the area under the ROC curve (AUC) for both the training and test data set. What does the difference in AUC between training and test tell you about your model, if anything?

7. Secondly, you will apply some of the techniques you have learned so far on the course to improve the model you built in step 4. You can make use of any of the variables supplied in the data set. Use AUC as your performance measure to determine which model is best.
   When doing this work, you should consider methods dealing with:

   a. Data preparation and validation.

   b. Variable selection.

   c. Model structure. Consider including interaction terms or building a segmented model.

   d. Testing. How will you use your data to ensure testing is performed correctly?

8. Write a report describing the methods you tried and the results you achieved. Report the model built in step 4 and the final model you tried that performed best and give an interpretation of it, especially in contrast to the model you built in step 4.

9. Notes:

   a. Your report should be between 5 and 10 pages.
   b. Include R code in an Appendix, not in the main body of the report.
   c. Submit your report through Blackboard.
   d. Marks will be given for good explanation of your work and for presentation.
   e. The emphasis is on the quality of your work, rather than quantity. You can achieve more marks by applying a few techniques well, rather than doing many poorly.

APPENDIX A

This data you will use in your project is constructed from a Lending Club data set of personal loans (www.lendingclub.com).

Each observation refers to a single personal loan originated in the USA and includes application variables as listed in the table below, along with outcome variables indicating default.

Variable	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
avg_cur_bal	Average current balance of all accounts
chargeoff_within_12_mths	Number of charge-offs within 12 months
def_flag	Indicates where loan defaults, with three outstanding missed payments, within the past 2 years
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
emp_length_p	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts

Variable	Description
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose_p	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
term	Loan term. The number of payments on the loan. Values are in months.
total_acc	The total number of credit lines currently in the borrower's credit file
total_rev_hi_lim	Total revolving high credit/credit limit
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

- Note that some of the descriptions of variables are ambiguous or unclear. However, this is the only information about the data we have.