Data dictionary is as follows (ignoring the first column which is id):

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 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.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

1. Read data 'default of credit card clients.xls' [downloaded and unzipped from LMS, module : linear models]. Use function read_excel from package readxl. skip first row while reading the data (use option skip=1)
2. Create dummies using function CreatDummies
3. Break the data into two parts (80/20)
4. Run a lm model first and drop variables one by one on the basis of vif cutoff 10
5. With **remaining** variables build logistic regression model with using function glm [family='binomial']. Drop variables on the basis of p-values.
6. Check the performance of the final model on the 20% data [auc score]