STATS 3001 Statistical Modelling III
Group Project
2019

This project is due by 23:59pm Tuesday 11 June 2019.

The team components should be submitted by email to `gary.glonek@adelaide.edu.au`. The individual components should be submitted on MyUni.

Introduction

The purpose of the project is to use statistical methods to develop a model for predicting default of credit card payment. Data are provided for 20,000 credit card customers in Taiwan and consists of a single binary response, indicating default in the October 2005 payments and 23 predictor variables. These data are your *training data.*

The task is to produce a predictor for default in the October payment. You can use any statistical method you believe provides the best result whether or not it is covered in this course.

The deliverables

Each team must submit, by the due date:

1. A function written in R that takes a $n \times 23$ matrix of predictor variables as its input and produces a n-dimensional vector of predictions with 0 for no default and 1 for default as its output. The function should be provided as a single text file of R code alled `prediction.r`.
2. A report of no more than 15 pages, documenting the predictor. It should contain a detailed description of the R function provided and documentation of the analysis that was performed to develop the function.

Separately, each student must submit:

1. A one-page essay reflecting on the structure of their group and the roles that each member played.
2. A percentage allocation of each team member's contribution the project.

The two individual components are to be submitted on MyUni by the due date and will not be available to other team members.

Assessment

Performance 60% A test set of 5,000 observations has been reserved to test the performance of each group's the predictor using the rate of correct classifications.

0/60 No working submission. That is, no submission at all or `R` function does not work.

10/60 `R` function works, but correct classification rate is no better than classifying all cases to the most common category.

30/60 `R` function works and correct classification rate is equivalent to a simple statistical model, such as logistic regression with only main effects.

30/60-60/60 The project that produces the best correct classification rate will receive 60/60. All other marks will by obtained be interpolation, based on the rate of correct classifications.

Report 30% The purpose of the report is to document the construction of your predictor. The primary criterion is that I should be able to recreate the steps to construct your predictor, based on the information in the report. The marks for performance of your predictor may not be awarded if an adequate description is not provided. In addition, the rationale for your analysis and any related diagnostic plots should be included.

Individual essay 10% This essay should be a reflective discussion of your experience working in a group.

About the data

The training data are available on MyUni in the file `train.csv`. The data consist of a binary response, Y, that indicates default ($Y = 1$) or no default ($Y = 0$) in the October 2005 payment and 23 predictor 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.

Further information and tips

The data are part of a larger data set available at `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`. The data for this project were chosen using a non-random method to select 25,000 records. 5,000 of those records were randomly selected to form the test set and the remaining 20,000 have been provided in the file `default.csv`. *Because the data for the project have not been randomly selected, the larger data set cannot be used to obtain a better model.* Your analyses should be performed solely on the data provided on MyUni.

This is a difficult prediction problem. If there was an easy way to identify future defaulters, they would have already been identified and not given credit cards. For this reason, even the best predictors may produce only a small improvement over the simple rule of always predicting the most common category. You should not be discouraged by this and, as explained above, it will not prevent you from getting a high mark for your project.

Beware of over-fitting. A simple approach would be to fit a very comprehensive model to the data provided. This might provide a very close fit to your data but the predictor you derive will perform badly on the test data. For this reason, you should consider a method such as cross-validation in developing your predictor. See for example, `https://en.wikipedia.org/wiki/Cross-validation_(statistics)` and the `caret` package in `R`.

⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆