Modelling competition

A copy of the report and the R markdown code has to be sent before the 20th of January at 20:00

1. Credit scoring

If banks give a credit to a client, they are interested in estimating the risk that the client will not pay back the credit as agreed by contract. The aim of credit-scoring systems is to model or predict the probability that a client with certain risk factors is to be considered as a potential risk. Credit scoring methods became standard tool of banks and other financial institutions, direct marketing retailers and advertising companies to estimate whether an applicant for credit/goods will pay back his liabilities.

The success of credit scoring in credit cards issuing was a significant sign for the banks to use scoring methods to other products like personal loans, mortgage loans, small business loans etc. However, commercial lending is more heterogeneous, its documentation is not standardized within or across institutions and thus the results are not so clear. The growth of direct marketing has led to the use of scorecards to improve the response rate to advertising campaigns in the 1990s.

Most of the problems one must face when using credit scoring are rather technical than theoretical nature. First of all, one should think of the data necessary to implement the scoring. It should include as many relevant factors as possible. It is a trade-off between expensive data and between low accuracy due to not enough information. Banks collect the data from their internal sources (from the applicants previous credit history), from external sources (questionnaires, interviews with the applicants) and from third parties. From the applicants'background the following information is usually collected: age, gender, marital status, education, job, income, lease rental charges, etc. The following questions from applicants credit history are especially interesting: has the applicant another credit?, what is the amount?, has the applicant ever delayed his payment?, etc. The variables entering the credit scoring procedures should be chosen carefully, as the amount of the data may be vast indeed and thus computationally problematic.

2. Description of the data

The Credit data set 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 and the definition of the variables are in the file `credit.xls`.

Take into account that you can define new variables from the ones in the dataset, re-code

them or transform them.

3. Task

Imagine that you work for a bank, and your task is to find the best model with the minimum number of variables that best predict the probability of being a good client and also that best classifies the individuals.

You will need to think about all the technical problems that can arise in regression and decide which variables should be treated as categorical. You also will probably wish to carry out some hypothesis tests as part of your work, check for interactions, transform or create new variables, etc...

3.1. Report

The written report of your project assignment should have, at least:

1. Model-building steps (exploratory analysis, model fitting, variable selection, etc.).
2. Best model/models selected (There are more than one good model, so give reasons to justify your election).
3. Predictive power of your model within sample (train dataset) and out of sample (test dataset)
4. Miss-classification rate
5. Other questions to be assessed:
   - What is the variable that contributes to the model?, what are the characteristics of the best/worst client?
   - Any other question that you consider interesting

3.2. Evaluation

The grade on the assignment will depend on three things:

1. **Clarity** The report must clearly indicate the steps you follow in your work, including which models are considered. Being able to communicate effectively is extremely important for a researcher. Excellent work is of no value if no other than the investigator can understand it.
2. **Content** The analysis should use the tools developed in the course in an appropriate and correct manner. The report should anticipate questions that a critical reader might ask.

3. **Presentation** The report should be **12 pages at most**, and should be easy to read, use tables and figures that are relevant and omit unnecessary graphs and/or code in the text.