QBUS2820
Predictive Analytics
Semester 2, 2017

Classification Project (Group Assignment 2 Task 1): Modelling Consumer Response to Marketing

1. Key information

Required submissions: Written report (by Turnitin submission) and Jupyter Notebook (through Ed).

Submission format: The format can be any accepted by Turnitin (one file).

Deadline: Friday November 3rd at 5pm.

Weight: 18 out of 100 marks in your final mark for the unit.

Groups: You can complete the assignment in groups of up to four students. There are no exceptions: if there are more than four you need to split the group. You will need to sign up to Blackboard groups for identification purposes.

Length: The main text of your report for this task should have a maximum of 20 pages. If you wish to include additional material, you can do so by creating an appendix. There is no page limit for the appendix. Keep in mind that making good use of your audience's time is an essential business skill. Every sentence, table and figure has to count. Extraneous and/or wrong material will reduce your mark no matter the quality of the assignment otherwise.

Marking and key rules:

- A separate rubric will indicate the marking criteria for the report.
- Carefully read the requirements for the assignment.
- Please follow any further instructions announced on Blackboard, particularly for submissions.
- You must use Python for the assignment. It is fine to use Excel for data manipulation, though this is neither efficient nor recommended.

- Failure to read information and follow instructions may lead to a loss of marks. It is your responsibility to be informed of the University of Sydney and Business School rules and guidelines for assignments, and follow them.

2. Problem description

This assignment is based on data from a clothing store chain that uses different marketing strategies to promote sales. Your task is to classify which customers will respond to direct mail marketing based on data collected for past customers.

The ultimate objective is to the increase the profitability of the store, so that the first step of the project is to build a cost-benefit table as below. The group should meet to decide how to fill in the blanks, and later justify the choices in the report. There is no unique right answer here – come up with numbers that you find reasonable.

Outcome	Classification	Actual response	Cost
True negative	Nonresponse	Nonresponse	_________
True positive	Response	Response	_________
False negative	Nonresponse	Response	_________
False positive	Response	Nonresponse	_________

The dataset is available for download on Blackboard.

3. Cross-Industry Standard Process for Data Mining (CRISP-DM)

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology that specifies a six-phase life cycle for data mining projects:

1. Business understanding.
2. Data understanding.
3. Data preparation.
4. Modelling.
5. Evaluation.
6. Deployment.

The data mining process is adaptive and iterative the sense that there are several dependencies between these phases, and the results from different steps may lead you to revisit earlier phases. The six steps above are similar to our discussion in the first lecture, but in the specific context of data mining.

As part of the assignment, you should familiarise yourself with the CRISP-DM methodology and apply this knowledge to the project. There are several resources on CRISP-DM available online, for example http://crisp-dm.eu/.

4. Report and requirements

- The purpose of your report is to summarise your solution to the business problem and the key insights from your analysis.
- The main report must address all the phases of the CRISP-DM methodology. The space is very limited, so that you may need to refer to the appendix for details.
- Your analysis should take both predictive power and interpretability into consideration (possibly through different methods).
- The model evaluation should be based on at least two substantively different models.
- The model evaluation should include a confidence interval for the expected gross profit per customer contacted based on each model.
- When splitting data, the random state needs to be the SID of one of the members in the group.

5. Dataset description

The response variable is the last column in the dataset. The other variables are the following.

- Customer ID.
- Zip code.
- Number of purchase visits.
- Total net sales.
- Average amount spent per visit.
- Amount spent for each of four different franchises (four variables).
- Amount spent in the past month, the past three months, and the past six months.
- Amount spent the same period last year.
- Gross margin percentage.
- Number of marketing promotions on file.
- Number of days the customer has been on file.
- Number of days between purchases.
- Markdown percentage on customer purchases.
- Number of different product classes purchased.
- Number of coupons used by the customer.
- Total number of individual items purchased by the customer.
- Number of stores the customer purchased at.
- Number of promotions mailed in the last year.
- Number of promotions responded to in the past year.

- Promotion response rate for the past year.
- Product uniformity (low score = diverse spending patterns).
- Lifetime average time between visits.
- Microvision lifestyle cluster type (market segmentation category defined by Claritas Demographics).
- Percent of returns.
- Flag: credit card user.
- Flag: valid phone number on file.
- Flag: web shopper.
- 15 variables providing the percentages spent by the customer on specific classes of clothing.