# Leveraging AI and Deep Learning Tools in Marketing (RSM8521) Assignment 3

This assignment’s purpose is to model customer data. You will be given a dataset to build a model where you will be graded in a *blind* test of your model’s performance. Additionally, you will be asked to create a report that describes the approach you took to modelling. Details of the problem are in the section below.

### **Problem**

In this problem you will be given a dataset containing customers, promotions given to them, and their transactions. Your task will be to predict which customers will respond to the given promotions. The dataset contains these tables:

* **transactions.csv.gz**: Transactions for all customers between 2012-03 and 2013-02
* **promos.csv.gz**: Metadata about each promotion
* **train\_history.csv.gz**: Promotions given to a subset of the customers during 2013-03 and whether or not they responded -- this will form your training data
* **test\_history.csv.gz:** Promotions given to a different subset of the customers during 2013-04 -- this will be your blind test data where you have to predict whether or not they will respond

All files are CSV format compressed with gzip. All data has been anonymized to preserve the originating data source. Please see the accompanying data dictionary spreadsheet for more details on the individual fields in each table.

The prediction output format has two columns: customer id (**id**) and your prediction as a probability between 0 and 1 of whether or not they will respond (**active**). It should contain every customer id listed in the test history. It will be evaluated using ROC AUC with the true labels in the test history.

An accompanying sample notebook gives Python code to:

* Read the files from your Google drive (you must download the data files and place them in your drive somewhere and the notebook will prompt you to give access to read them via an authorization code)
* Extract some basic RFM features
* Generate a train and test set
* Fit a simple random forest model
* Evaluate on the test set
* Generate predictions and allow you to download them

Feel free to use this notebook as a starting point and modify freely.

### **Additional Rules**

* You may not use any external data sources
* All code must fit within a single .ipynb notebook and be executable on Google Colab
* Notebook must take less than 10 minutes to train the model from scratch and generate the predictions (hyperparameters can be fixed and tuned prior to submission)
* You may use any public libraries that are available or can be installed on Google Colab (but library installation time is included in the 10 minute limit above)
* Your submitted notebook must *exactly* generate your submitted CSV file

### **Hints**

* Engineering good features are likely to be more important than selecting the perfect model
* Generalization is going to be the biggest challenge, think about ways to robustly evaluate your model as well as techniques you can apply to help with it
* Make sure your training data does not leak any data from the validation/test set
* When making the final prediction, be sure to use all your data to train the model
* The transactions dataset is of moderate size so be careful when doing inefficient transformations, which will slow you down and eat into your 10 min runtime budget
* The sample notebook does *not* use the promotion metadata nor does it use any details about what items were purchased, how might they be useful?

### **Report**

The report will simulate writing a technical report that you might be asked to write on the job. The report should be a maximum of 4 pages containing the following sections:

* Brief introduction summarizing the problem and the main points about the dataset.
* Describe your approach in prose (*not code*), include things that both worked and did not work. Be sure to include details on your features, model, and any additional “tricks” that you implemented.
* Describe any non-obvious code-level implementation details that you had to use (can use bullet points and code snippets).
* Describe your experimental setup used to validate the model and report the relevant metrics.
* Future directions that you would take if you had more time and also include what you would do differently next time (if anything).

Be sure to include references and present the material in a professional manner (e.g. proper formatting, grammar, spelling etc.).

### **Grading**

You will be graded out of a total of 55 marks as follows:

* **Blind Test Performance**: 20 marks
  + AUC 0.0 - 0.55: 0 marks
  + AUC 0.55 - 0.60: 5 marks
  + AUC 0.60 - 0.65: 10 marks
  + AUC 0.65 - 0.70: 15 marks
  + AUC 0.70 - 0.75: 20 marks
  + AUC 0.75+:: 25 marks (bonus marks)
  + AUC 0.80+: 30 marks (bonus marks)
* **Report**: 35 marks
  + 5 marks for each of the 5 major sections listed above
  + 10 marks for structure including references, grammar, formatting etc.

If your CSV file does not conform to the specification or your notebook is not able to run, the grader will deduct marks at their discretion, which may include giving a zero on the relevant portion if the submission is unsalvageable.

### **Submission**

Please submit individual assignments via Quercus with the following:

* A PDF document of your report
* A .ipynb notebook of your model training and prediction code
* A CSV file in the output format specified above that contains your predictions for the test history

### **Collaboration & Academic Honesty**

You may collaborate with others to try to solve these problems, however, this should be limited to ***only verbal discussions***. You should indicate any collaborators on your submission. You may also use any references on the internet but you should clearly indicate your sources and which parts you used.

You may **NOT** share code or written answers with any classmates, this would be considered cheating. You should be attempting to solve the problem by typing, debugging and running the code yourself. This is the only way you will actually learn the material!