# DAM Assignment 2 – Credit Card Default Model

## Background

This task is to predict which customers will default on their credit card repayments next month. The data set is based on the publicly available credit card default data set from the UCI Machine Learning Repository. Details of the original data are here:

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

A popular kaggle competition has also been run based on this data. Details are here:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

In this task, some new data has become available. This new data has distorted the original relationship between the default target and the predictors. Also, Principal Components Analysis has been applied to the 6 original variables for the history of past payment statuses, and also to the 12 original variables for the past bill amounts and the past payment amounts. This reduces the first 6 variables to 3 principal components, and the second set of 12 variables to 7 principal components, while retaining over 90% of the variation.

You are free to explore previous work that has been completed on this data, from either the original data set or the kaggle competition. However, caution is advised in using other people's solutions since this new data requires some new interpretation and a newly trained model.

Here is a description of the data:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_PC1, PAY_PC2, PAY_PC3: First three Principal Components of repayment status from April to September, 2005
- AMT_PC1, AMT_PC2, AMT_PC3, AMT_PC4, AMT_PC5, AMT_PC6, AMT_PC7: First seven Principal Components of the bill statement amount and the amount of previous payments from April to September, 2005
- default: Default payment next month (1=yes, 0=no)

You can work in teams of your choice for this assignment.

There are two key deliverables for this assignment, Part A and Part B.

## Part A – Modelling (GROUP ASSIGNMENT)

Create a model to predict which customers are likely to default on their credit card repayments next month. The data and submission process are managed via a kaggle competition. There will also be a live leaderboard. The link to the competition is here:

https://www.kaggle.com/t/37f1f6701a3049bfb27dc7cd5b80971f
https://www.kaggle.com/c/damat2-credit-spring-2018

The former is the privacy link required to ensure the competition is not open to the public

There are two data sets, one training and one validation:

- AT2_credit_train_STUDENT.csv
- AT2_credit_test_STUDENT.csv

The performance of your model will be evaluated using the AUC measure (Area Under the ROC Curve) for binary classification models on the validation data set. One part of the validation data is public and will be made visible once you submit your predictions, but a second part is private and will be withheld until the assignment is finished.

You can see the AT2_credit_sample_UPLOAD.csv file for how to format your submissions. Note you are limited to 2 submissions per day to prevent gaming of the test set so make sure you put your best feet (and by that we mean models) forward!

**Deliverables:**

The assessment brief for DAM outlines that a report is also due for this component as well as a statement of contributions. Therefore there are three deliverables for this task (these only need to be submitted once per team).

1. Complete, commented R-code onto Canvas
2. Report following the CRISP-DM framework including:
    a. The business problem
    b. The available data
    c. Your data preparation process
    d. Any particular insights you discovered about the data
    e. Details of your model training, including the assumptions you made with a rationale for why you adopted this process
    f. Your evaluation methodology
    g. Preliminary results (kaggle public evaluation measures)
    h. Consideration of ethical issues
3. A statement outlining contributions of each team member for this assignment

## Part B – Management Presentation (INDIVIDUAL ASSIGNMENT)

Each member of the team is required to submit a management presentation on your approach. You should reduce and *alter* your report to align to this different audience. It is **crucial** to ensure that your presentation is appropriate for senior management, that are largely non-technical in background.

Your presentation should be short and concise, no more than 10 slides.