# Overview

In this assessment, you will implement and compare three machine learning algorithms on a real data. The assessment addresses the following learning outcome(s):

* identifying and translating a data science problem into a supervised learning problem;
* identifying appropriate lasso-regression, tree-based methods, and support vector classification for descriptive problems;
* application of lasso-regression, support vector classifier and tree-based methods to a dataset using the computer language R and the software environment RStudio.

# Submission

You will need to submit the following:

A PDF/Word file clearly shows the question, the associated answers, any relevant R outputs, analyses and discussions. Please attach R code script in Appendix.

**Rmarkdown/R** script file to reproduce your work.

**A word on plagiarism:**

Plagiarism is the act of using another’s words, works or ideas from any source as one’s own, this includes the use of large language model, such as ChatGPT. Business work containing plagiarised material will be subject to formal penalty processes.

# Problem

## Background on Credit Card Dataset

The data, **“CreditCard Data.xls”**, is based on Yeh and Lien (2009). The data contains 30,000 observations and 23 explanatory variables. The response variable, Y, is a binary variable where “1” refers to *default payment* and “0” implies *non-default payment*. The description of 23 explanatory variables is as follows:

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 (0 = unknown; 1 = graduate school; 2 = university; 3 = high school; 4 = others; 5 = unknown; 6 = unknown).

X4: Marital status (0 = unknown; 1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. The data was 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: -2= no consumption, -1=pay duly, 0 = the use of revolving credit; 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.

## Assessment Tasks

### Data

(a) Select a random sample of 70% of the full dataset as the training data, retain the rest as test data. Provide the code and print out the dimensions of the training data.

### Lasso regression

1. Use lasso-regression to find the best model which classifies credible and non-credible clients. Specify any underlying assumptions. Justify your model choice as well as hyper-parameters which are required to be specified in R.
2. Display model summary and discuss the relationship between the response variable versus selected features.
3. Evaluate the performance of the algorithm on the training data and comment on the results.

### Tree-Based Algorithms

1. Use an appropriate tree-based algorithm to classify credible and non-credible clients. Specify any underlying assumptions. Justify your model choice as well as hyperparameters which are required to be specified in R.
2. Display model summary and discuss the relationship between the response variable versus selected features.
3. Evaluate the performance of the algorithm on the training data and comment on the results.

### Support vector classifier

1. Use an appropriate support vector classifier to classify the credible and non-credible clients. Justify your model choice as well as hyper-parameters which are required to be specified in R.
2. Display model summary and discuss the relationship between the response variable versus selected features.
3. Evaluate the performance of the algorithm on the training data and comment on the results.

### Prediction

Apply your optimal models identified in section, and make prediction on the test data. Evaluate the performance of the algorithms on test data. Which models do you prefer? Are there any suggestions to further improve the performance of the algorithms?

Justify your answers.

# References

Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480.