The proposal should address the following questions:

* What is the problem you want to solve?
  + How do we effectively predict credit card payment default?
  + How does the probability of default payment vary by categories of different demographic variables?
  + Which variables are the strongest predictors of default payment?
* Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?
  + *Do some research on credit card companies, how they try to calculate default probabilities, and how they use those numbers*
    - In other words:
      * What decisions credit card companies make with their models
      * Value of an improved decision
  + My client is the credit card company that is trying to predict whether customers will default on their next payment
  + They care about this problem, because knowing with greater certainty who will default on their payments allows them to
    - A) Extend credit to those who are (very) likely to pay on time
    - B) Extend interventions to those who are likely to miss their payments to prioritize their payments ahead of other expenses and have them pay on time
  + If both of these things are done successfully, the company will save significant money
  + Could be generalized to other problems like bankruptcy?
* What data are you using? How will you acquire the data?
  + Using a dataset that “contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.”
    - Found on Kaggle: <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home>
    - Originally from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#
* Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
  + This project will implement 6 common data mining techniques, as outlined in the paper “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients” [1]. They are:
    - K-nearest neighbor classifiers (KNN)
    - Logistic regression (LR)
    - Discriminant analysis (DA)
    - Naı¨ve Bayesian classifier (NB)
    - Artificial neural networks (ANNs)
    - Classification trees (CTs)
  + In particular, this project will attempt to recreate the results reached by that paper and note if any differences of result are reached (and explore them)
* What are your deliverables? Typically, this includes code, a paper, or a slide deck.
  + Deliverables for this project include project code, a project paper or report, and a presentation of the project findings in a slide deck.

[1] Yeh, I. C., & 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.

# Capstone 1: Default of Credit Card Clients Dataset

## Default Payments of Credit Card Clients in Taiwan from 2005

### Dataset Description

* information on
  + default payments
    - default.payment.next.month: Default payment (1=yes, 0=no)
  + demographic factors
    - X2: Gender (1 = male; 2 = female)
    - X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
    - X4: Marital status (1 = married; 2 = single; 3 = others).
    - X5: Age (year)
  + credit data
    - X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
  + history of payment
    - X6 - X11: History of past payment.
    - X18-X23: Amount of previous payment (NT dollar)
  + bill statements of credit card clients in Taiwan
    - X12-X17: Amount of bill statement (NT dollar)
* from April 2005 to September 2005

### Task/Problem/Question

1. How does the probability of default payment vary by categories of different demographic variables?
2. Which variables are the strongest predictors of default payment?