# Write up for HW2

## 1. Goal

The goal of this assignment is to build a simple, modular, extensible, machine learning pipeline in Python.

## 2. Problem to solve

The task here is to predict who will experience financial distress in the next two years. The outcome variable (label) in the data is SeriousDlqin2yrs.

## 3. Pipeline:

The pipeline should have functions that can do the following tasks:

1. Read/Load Data

   The read_data function can read csv, xls, and json type of file. The assuming file type for this assignment is csv.

2. Explore Data

   The data_overview function can give people overall statistical summary of all the feature variables and the outcome variable. The make_graph function can generate simple histogram for each feature variable, which enable people to better understand the distribution of the features.

3. Pre-Process and Clean Data

   The fill_null function can fill the missing values for the desired feature variable. Depending on the situation, there are two ways here to fill the null, one is to median, and the other is to use average. Zero can also be chosen in some other situations.

4. Generate Features/Predictors

   Two functions are created here. One is to discretize a continuous variable, and the other is to take a categorical variable and create binary/dummy variables from it. The two functions are applied to the following feature variables: number of dependent, and monthly income.

5. Build Machine Learning Classifier

   Five models are chosen here, including logistic regression, k-neighborhood classifier, decision tree classifier, random forest classifier, and gradient boosting classifier.

6. Evaluate Classifier

   Accuracy, recall, and precision scores are used here to roughly evaluate the five models. The accuracy scores for all of them are around ninety percent. For this particular question related to financial distress, we should focus more on the false negatives rather than false positives.

## 4. Future Improvement

More detailed analysis should be conducted on the feature variables in order to find out the variables that really matter to the outcome variable. For the evaluation part, more strict and rigorous methods should be applied to figure out the most suitable model to use and the real accuracy of the applied test model.