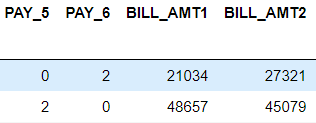
## Introduction

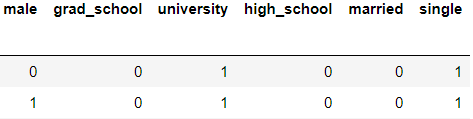
The data chosen in this project is called “Default of Credit Card Clients Dataset”. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. This is the link to the dataset: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. The aim of this project is to construct a predictive model using various machine learning algorithms.

## Data Cleaning

1. 

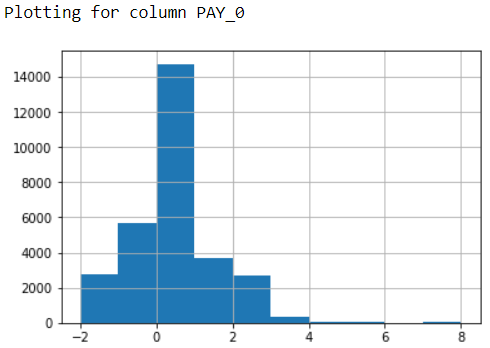
*Figure 1.1 shows that some of the features have very different scales*

When features are highly varying in range, the features with high magnitudes will have bigger effect towards the distance calculations as compared to features with low magnitudes. This will affect the learning performance of classifier, and classification accuracy and precision. Therefore, the RobustScaler is used to scale the features to a common range. RobustScaler is chosen because it uses statistics that are robust to outliers so the model training process will not be affected by outliers yielding a more accurate and precise result.

1. 

*Figure 1.2 shows the one hot encoding of three categorical attributes*

The SEX, EDUCATION and MARRIAGE are categorical data with their values represented by numerical values such as 1, 2, 3 and so on. This might confuse the machine learning algorithms into thinking that these attributes have some kind of order or hierarchy and take into account the magnitudes of the values in the training process. As a result, the classification result might not be correct and therefore, these three attributes are one hot encoded to avoid this issue.

1. 

*Figure 1.3 shows the distribution of column PAY\_0*

Figure 1.3 shows that the PAY\_0 column contains undocumented categories which are -2 and 0. The same goes to the other PAY\_n columns. According to the documentation, the PAY\_n variables are just variables represent the number of months (payment delay) and -1 is used to indicate "pay duly". Therefore, we can infer that PAY\_n features represent payment delayed for 0 month i.e. “pay duly” if it is less than or equal to 0. To solve this issue, all values of PAY\_n features which are less than 0 are modified to belong to the category 0.

## Hyperparameter Tuning

1. Logistic Regression