Credit Card Client - Default Data Set

Abstract: This research aimed at the case of customersâ€™ default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.

1. **Data (Attribute) Characteristics -** MultiVariate **(**Integer, Real**);**
2. **Associated Tasks :** Classification Supervised Learning.
3. **Number of Instances & Attributes :** 30000 & 24
4. **Area :** Business - Banking (Credit Card - Default)

Data Set Information:

This research is aimed at the case of customers' default payments in Taiwan and compares the **predictive accuracy of probability of default** among **six data mining methods**. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel "**Sorting Smoothing Method"** to estimate the real probability of default.

With the real probability of default as the **response variable** (**y**), and the predictive probability of default as the **independent variable** (**X**), the simple linear regression result ($y = beta_0 + beta_1 X_i$) shows that the forecasting model produced by an artificial neural network has the highest coefficient of determination; its regression intercept ($beta_0$) is close to zero, and regression coefficient ($beta_1$) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

Attribute Information: This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

- **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 (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- **X4** : Marital status (1 = married; 2 = single; 3 = others).
- **X5** : Age (year).
- **X6** - **X11** : History of past payment. We 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: -1 = pay duly; 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.