TAIWAN CREDIT DEFAULT

Group No.: Group 23

Student Names: Devarsh Shah and Gaurav Handa

I. Background and Introduction

Money lending which is claimed to be world's one of the oldest professions and thus its analysis of credit risk is equally important. Credit default system has been prevalent in financial institutions for determining which attributes are useful in determining the probability of a customer being a default or not. It is important to understand how these attributes relate together is a systematic and logical manner with considerate statistical proof to back it up rather than human judgement.

II. Data Collection

The given data taken in October 2005 taken from a bank in Taiwan has 30000 records with 23 attributes and with response variable – Default payment (Yes = 1, No = 0).

The given 23 attributes are explained:

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.

III. The problem

There are many different data mining techniques (K nearest neighbor, logistic regression, Classification trees, Naïve Bayes) which can be used on this given problem, but it is difficult to determine which can give more accurate probability of default. The problem is to find which type of data mining technique is most relevant to this given problem giving highest levels accuracy.

IV. Possible Solution

Possible solution is to break the dataset into training data, validation data and test data and apply them on given data mining techniques and compare the results of proportion of predictive accuracy, error rates to determine which model is better fit for the given problem and accurately determine the probability of default.