Prediction of potential credit default accounts

Keywords: *Classification, Credit Default, Prediction*

Introduction

One of the key decisions financial institutions have to make is to decide whether to sustain a loan to a customer or not. This decision reduces to a binary classification problem which aims at distinguishing decent payers from bad ones. Until recently, this distinction was made using a rudimentary approach by simply inspecting the financial situation and history of the applicant. It is then decided upon the creditworthiness of the applicant, using all possible relevant information, and intentions.

The revolution of data technology has facilitated financial institutions' ability to store all information regarding the characteristics and repayment behaviour of credit applicants. This has motivated the need to automate credit scoring by using statistical or machine learning algorithms.

Historically, many statistical methods, including discriminant analysis, logistic regression, Bayes classifier, and nearest neighbor, have been used to develop models of risk prediction. With the evolution of artificial intelligence and machine learning, artificial neural networks and classification trees were also employed to forecast payment default.

Objectives

The purpose of this project is to use personal and financial information such as customer transaction and repayment records to predict individual customers' behavior concerning loan repayment using machine learning algorithms, which will allow banks to make important decisions, concerning for example credit extension, therefore reducing risk and damage.

Our aim is to work on the payment data in October, 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank

The input of our project is considered to be the dataset (Fig.1) consisting of many variables such as clients' genders, education, marital status, etc.

Methods

First, we explore the data: read it, complete it if there are missing values, and then look to apprehend the data by looking at the correlations and the most important variables using PCA/SVD. We will also look at proposing some new features if necessary.

Then, we will separate our data into training set and validation set using a 80% / 20% method or and K-fold for cross-validation.

So the goal is to forecast the response variable : default payment, in order to do that, and since the response variable takes values 0 or 1, we will train our function on the dataset, using binary classification algorithms such as : logistic regression, decision tree, Random Forest, Gradient Tree Boosting, Adaboost, XGBoost and so on to predict the binary label: default payment.

Model Evaluation

After training our model using one of the classifiers, we will use the AUC method or Confusion matrix method to determine the performance of our model. This evaluation will be repeated for each classifier algorithm so we can compare the performance between them.

References:

- **Desai, V. S., Crook, J. N., & Overstreet, G. A. A.** (1996). *Comparison of neural networks and linear scoring models in the credit union environment*. European Journal of Operational Research, pages 24–37.
- **Cheng Yeh, Che-hui Lien** (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*. Expert Systems with Applications, Volume 36, Issue 2 Part 1, pages 2473-2480.
- **Hand, D. J., & Henley, W. E.** (1997). *Statistical classification methods in consumer credit scoring: A review*. Journal of the Royal Statistical Society, Series A – Statistics in Society, pages 523–541.
- **Baesens, B., Setiono, R., Mues, C., & Vanthienen, J.** (2003). *Using neural network rule extraction and decision tables for credit-risk evaluation.* Management Science, pages 312–329.
- **Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, Lyn C.Thomas** (2015). *Benchmarking state-of-the-art classification algorithms for credit scoring*. Journal of the Operational Research Society, Volume 247, Issue 1, pages 124-136

X1 : Amount of the given credit (NT dollar)	Individual consumer credit & 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	X6 = Repayment status in September 2005 : (-1=pay duly; n=payment delay for n (>0) months starting from September 2005) X11 = Repayment status in April 2005
X12 – X17: Amount of bill statement (NT dollar)	X12 = amount of bill statement in September 2005 ..., X17 = amount of bill statement in April, 2005
X18 – X23: Amount of previous payment	X18 = amount paid in September, 2005... X23 = amount paid in April, 2005
X24 : Default Payment	Yes=1 , No=0

Fig.1 Dataset