BUSN 41201 Final Project

Arthur Cheib, Yu-Wei Chen, Simone Zhang, Sheng-Hau Peng, Shu-Hsiang Wang

2023-05-30

Contents

Summary

Introduction

Dataset Information

Source:

Name: I-Cheng Yeh

Email addresses: (1) icyeh '@' chu.edu.tw (2) 140910 '@' mail.tku.edu.tw

Institutions: (1) Department of Information Management, Chung Hua University, Taiwan. (2) Department of Civil Engineering, Tamkang University, Taiwan. Other contact information: 886-2-26215656 ext. 3181

Data Set Information:

This research 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 ($Y = A + BX$) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) 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.

Exploratory Data Analysis

```
library(readxl)
data = read_excel("default_of_credit_card_clients.xls", skip = 1)
```

```
# eda
summary(data)
```

```
##        ID          LIMIT_BAL            SEX          EDUCATION
##  Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000
##  1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000
##  Median :15000   Median : 140000   Median :2.000   Median :2.000
##  Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853
##  3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000
##     MARRIAGE          AGE            PAY_0             PAY_2        
##  Min.   :0.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000  
##  1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000  
##  Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000  
##  Mean   :1.552   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338  
##  3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
##  Max.   :3.000   Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000  
##      PAY_3             PAY_4             PAY_5             PAY_6        
##  Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000  
##  1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000  
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000  
##  Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911  
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
##  Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000  
##    BILL_AMT1         BILL_AMT2        BILL_AMT3         BILL_AMT4      
##  Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000  
##  1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327  
##  Median :  22382   Median : 21200   Median :  20089   Median :  19052  
##  Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263  
##  3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506  
##  Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586  
##    BILL_AMT5        BILL_AMT6          PAY_AMT1         PAY_AMT2      
##  Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0  
##  1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833  
##  Median : 18105   Median :  17071   Median :  2100   Median :   2009  
##  Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921  
##  3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000  
##  Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259  
##     PAY_AMT3         PAY_AMT4         PAY_AMT5           PAY_AMT6       
##  Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0  
##  1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8  
##  Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0  
##  Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5  
##  3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0  
##  Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0  
##  default payment next month
##  Min.   :0.0000            
##  1st Qu.:0.0000            
##  Median :0.0000            
##  Mean   :0.2212            
##  3rd Qu.:0.0000            
##  Max.   :1.0000            
```

```r
colnames(data)[25] = "default"
```

Questions that We Want to Solve

Model Building

Regression

Logistics Regression on Default

```
# logistic_fit = glm(default~ data[,2:24], data = data, family = binomial)
```

LASSO

Model Selection

Principle Components Analysis

Classification

KNN

SVM

Clustering

K Means

Hierarchical Clustering

Machine Learning

Artificial Neural Network

Conclusion

Potential Future Research

Appendix