信用卡用戶資料分析

壹、研究問題與研究動機

一、研究動機

台灣經濟社會環境變遷，人們對貨幣觀念逐漸轉型，因此信用卡需求量與日俱增。1996 年平均每人持卡數約 0.5 張，與美國的 4.5 張仍有一段差距。隨著全球金融市場之開放，開發中國家信用卡之服務趨勢明顯。

信用風險一直是金融機構關注的層面，若能從信用卡用戶的基本資料以及過往的信用卡使用情形預測該用戶是否屬於高風險客戶，進一步做出調降信用額度或其他措施，將可能降低部分風險和損失。另外，預測的正確率十分重要，將高風險用戶誤判為低風險會造成極大的損失；把低風險用戶誤判為高風險用戶，很可能用戶會因額度或福利遭限縮而與其他金融機構合作，造成企業客戶的流失。因此本研究探討高信用風險用戶的預測，以及著重模型預測的準確。

二、研究問題

問題一、與信用卡用戶被停卡機率相關的因素

問題二、哪些特徵的用戶被停卡的可能性較高

問題三、識別信用差、未來極可能付不出卡債的用戶

問題四、預測正確率較高的模型

貳、資料來源與資料陳述

一、資料來源: Kaggle 數據平台

二、資料陳述: 資料集--Default of Credit Card Clients Dataset 涵蓋 2005 年 4 月至 9 月台灣地區信用卡用戶基本資料、延遲繳付紀錄、歷史繳付金額、是否被停卡，共 85 個變數、3 萬筆樣本資訊。

參、分析

一、模型與分析方法

邏輯斯迴歸[逐步法] (Logistic Regression)

決策樹 (Decision Tree)

支持向量機（SVM）

樸素貝葉斯 (Naive Bayes)

K 鄰近算法（KNN）

二、分析過程

1. 以逐步刪除法配適出具有統計顯著性的邏輯斯迴歸模型，找出與停卡機率具顯著影響的變數。
2. 從模型中的估計參數觀察哪些特徵的用戶被停卡的可能性較高，高多少。
3. 本研究使用內部樣本模型測試，以訓練集建構的逐步邏輯斯迴歸模型預測測試資料集的用戶被停卡與否。使用 ROC (receiver operating characteristic) 圖形的 AUC 大小判斷是否模型預測優於隨機、是否可信，且可得此模型正確預測率。

模型可信下，藉由獲得信用卡客戶之基本資料、遲付紀錄，可預測該客戶被停卡機率，此時衍生出另一問題—「預測停卡機率多高時，要將客戶識別為信用差、未來極可能付不出卡債的用戶?」

在此最關心的是預測某用戶會被停卡，而未來真的被停卡，因此在給定此正確預測率最高的情況下，求得一停卡機率值作為將客戶歸屬為付不出卡債的分群界線，以此界線找出未來可能信用破產的客戶名單，先行列為警惕對象，避免貸方損失過多。

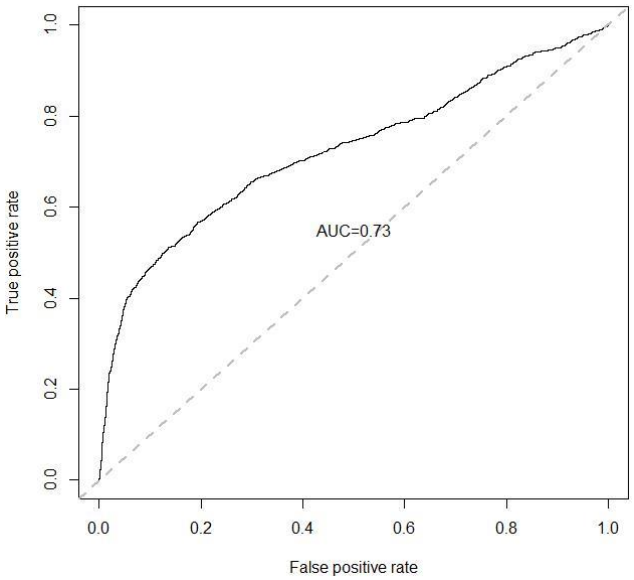
4. 探討各模型的正確預測率。

肆、結論

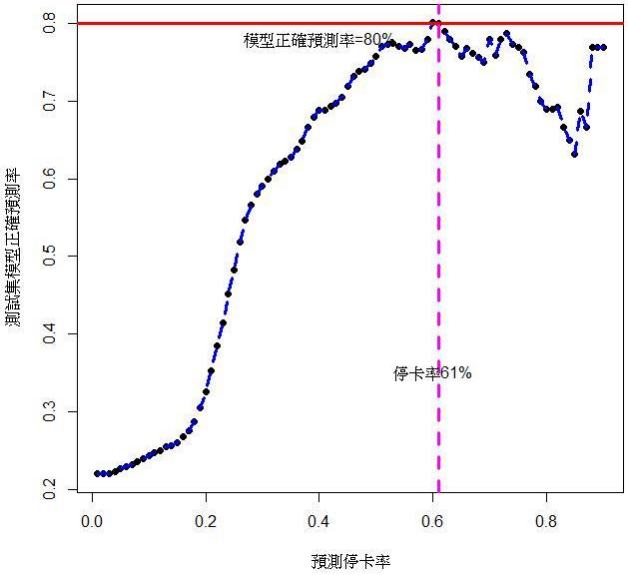
1. 停卡機率較高的用戶特徵-- 男性、教育程度高、已婚、年紀大、有過拖延付款記錄的人。
2. 根據 ROC 圖，邏輯斯迴歸模型 AUC 值: 0.73，優於隨機預測。*[附錄、圖一]*
3. 收集到信用卡用戶資料後，藉由模型預測該用戶被停卡的風險(機率值)，當風險高於 61%，列為潛在信用破產客戶，邏輯斯迴歸模型正確預測率達最高(80%)。*[附錄、圖二]*
4. 測試集中大約有 2.3%的用戶將被列為潛在信用破產客戶，可進一步追蹤此名單，降低信貸風險。
5. 比較不同模型被停卡的正確預測率，以 SVM、Naïve Bayes 三者為優，準確率 8 成 2，邏輯斯、KNN*[附錄、圖三]*次之，有 8 成左右的正確率，CART 正確率則僅 7 成。

附錄、

圖一、ROC 圖



圖二、測試集模型正確預測率_預測停卡率



圖三、KNN_最佳 K 值