(統計軟體與資料分析)期中報告

運用決策數進行信用評估

國立金門大學

資訊工程系

組長　魏仲彥

指導教授: 張伯銀教授

目錄

(一)前言

最近疫情盛行，消費型態也隨之改變，在家裡使用信用卡消費成為了不可或缺的方法之一，但也因如此，許多人成了月光族，永遠都在還債，甚至還不出錢來，為了避免這種狀況發生，本報告會以過去的信用卡資料建立模型，用來推測一個人否適合可以使用信用卡借款。本報告主要以監督式學習來建立模型，主要以 CART、KNN 和 SVM 分別對不同類型的人做分析。有別於傳統使用數學式判斷資料的方式。本專題會先講述傳統的信用分析模型，再慢慢帶入到監督式學習的模型，模型資料蒐集完後，再進行模型指標分析，以評估各方法性能。

(二)文獻探討

信用分析模型概述

信用分析模型是準確評估對象的信用等級和風險級別的關鍵技術，企業所在行業不同和客戶群的差異決定了信用分析模型設計的相對獨特性。Credit Risk(信用風險)，代表違約的程度與金額的多寡;而 Default Risk(違約風險)則代表僅代表對方無法償還某一期原先約定的本息，只有是或否的答案。

信用模型的類別

傳統的信用模型分為兩種預測模型和管理模型，其中，Z 計分模型、巴薩利模型式、營運資產模型和特徵分析模型是傳統分析方法中最常見的方式。

Z 計分模型

適用於大的集團公司，使用稅前利潤除以平均流動負債，來衡量公司業績;流動資產除以負債組數以及流動負債除以總資產數來衡量公司的債股比率;以現金交易間格期來衡量公司在收入狀態下可維持業務的時間長短，透過關鍵的財務比率來預測公司破產的可能性。

巴利薩模型

巴利薩模型是比 Z 計分法更普遍的應用，使用稅前利潤加上折舊加上遞延稅除以流動負債來衡量公司業績;稅前利潤除以營運資本來衡量營運資本回報率;股東利益除以流動負債來衡量股東權益對流動負債的保障程度;有形資產淨值除以負債總額衡量扣除無形資產對債務的保障程度;營運資本除以總資本來衡量流

動性,以上總和便是巴利薩模型的最終指數。低指數或負數均表明公司前景不妙。巴利薩模型最大優點在於易於计算,同時,它還能衡量公司實力大小,廣泛適用於各種行業。

營運資產模型

自1981年起在國外開始應用,該模型的計算分兩個步驟:一、營運資產計算:該模型首先提出考察的指標是營運資產,經此作為衡量客戶規模的尺度,這一指標與銷售營業額無關,只跟客戶的淨流動資產和帳面價值有關。營運資產的計算公式是:營運資產等於營運資本加上淨資產,其中營運資本等於流動資產減掉流動負債淨資產,即為企業自有資本或股東權益。二、資產負債表比率計算。營運資產模型考慮如下比率:(1)流動比率(流動資產除以流動負債); (2)速動比率(流動資產減掉存貨除以流動負債); (3)短期債務淨資產比率(流動負債除以淨資產); (4)債務淨資產比率(負債總額除以淨資產),評估值為(1)+ (2)- (3)- (4) ,(1)和(2)衡量公司的資產流動性;(3)和(4)衡量公司的資本結構。評估值綜合考慮了資產流動性和負債水平兩個最能反映公司償債能力的因素。評估值越大,表示公司的財務狀況越好,風險越小。以上傳統方式大多是對於公司進行信用評估,對於數據較少,或是個人通常都不適用,因此本專題將以監督式學習的方法對數據進行信用評估。

(三)研究方法:

本專題一共使用了三種模型進行資料模型建立,使用 KNN、SVM、CART 進行資料分析,使用 UCI Machine Learning 資料集進行數據分析,資料集輸入如下列表格呈現

	變量敘述	單位
x1	LIMIT_BAL。根據家庭狀況,給定信用額度	int
x2	SEX。1=男生; 2=女生	int
x3	EDUCATION。教育程度。1=研究生; 2=大學生; 3=高中; 4=其他	int
x4	MARRIAGE。婚姻狀況。1=已婚; 2=單身; 3=其他	int
x5	AGE。年齡(年)	int
x6~x11	PAY。過去的付款歷史,每月記一次。-2=沒有借款; -1=按時還款; 1欠款一個月; 2欠款量個月; .3=欠款3個月... 以此類推。	int
x12~x17	BILL_AMT。過去的帳單金額	int
x18~x23	PAY_AMT。上一次支付的金額	int
y	default payment next month (目標值)。1是可以還清,0是欠款	bool

KNN

KNN 方法主要靠周圍有限的鄰近的樣本，而不是靠判別類域的方法來確定所屬的類別，因此對於資料重疊較多的樣本集來說，KNN 方法較其他方法更為適合，所以 KNN 模型適合用於處裡資料種類偏向單一的客戶。
KNN 模型中，資料集使用 2015 年 4 月到 9 月(x6~x11)作為數入資料，y 作為預測值。根據還款狀況，判斷信用卡使用者使否可以繼續借款。

SVM

SVM 可以處理多維的數據，也可以同時處理線性和非線性的問題，其他的模型對於線性和非線性同時處理比較弱，如果遇到信息複雜的客戶，可以使用 SVM 當作信用評估模型。
SVM 模型中，資料集使用 2015 年 4 月到 9 月(x6~x11)作為數入資料，y 作為預測值。SVM 模型的分類準確率比 KNN 還要高，比 KNN 還更適合預測 y

CART

使用決策樹裡面的 CART 做為模型，可以針對許多不相關的特徵進行數據處理，而且 CART 是一個白盒模型，易於推導和觀察，不論是處理多種不一樣的數據還是單一的都可以有良好的結果，而且 CART 模型效率高建構，只需要一次建構，就可以反覆使用。
CART 模型中，資料集採用家庭狀況(x1)、性別(x2)、教育程度(x3)、婚姻狀況(x4)、年齡(x5)、還款狀況(x6~x23)作為數入資料，y 作為預測值，使用決策樹分析，可以讓預測值具有高度的可信性。

(四)預期結果

針對信用卡使用者數據進行模型建立，下面是根據 UCI Machine Learning 資料集所得到的模型結果。

資料分析評估

使用 KNN 對資料進行處理，如果使用原始的 23 欄位(6 種類型)，模型的指標是 0.751，而如果是使用單一種類進行資料評估(x6 ~ x11)，模型的指標就會上升到 0.811，而 SVM 的準確度則是從 0.779 上升到 0.820。 使用 SVM 對數據進行處理雖然可以處理複雜資料，對於單種類或複雜的處理也十分在行，準確率也比

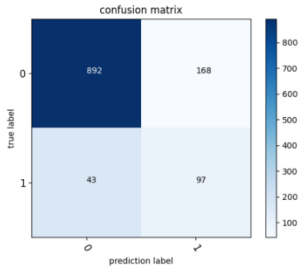
KNN 還要高，但運算量巨大，故下面使用決策樹，針對資料進行分類。

使用 CART 對資料進行分析，輸入 x1~x23，使用最大深度 3 進行數據分析，準確度可以達到 0.823，在模型中，可以很明顯的了解，信用卡使用者可不可以還款，大多都是從信用卡付款狀況可以做判斷，而跟家庭狀況、性別、信用額度和學歷沒有太大的關係。

模型指標分析

混淆矩陣

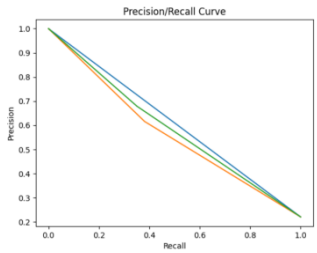
使用混淆矩陣預測結果(可以還錢，還不了錢)，從資料當中隨機選取數據，並使用 CART 模型預測數據，把相關的資訊呈現成混淆矩陣，可以從下圖中看到準確率為 0.8242（ 892 + 97 / (892+97+168+43) = 0.8242)。
TN 代表預測資料正確，實際資料也正確(左上)；TP 代表預測資料正確，實際資料不正確(右上)；FN 代表預測資料不正確，實際資料正確(左下)；TP 代表預測資料不正確，實際資料也不正確(右下)。



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

PR 曲線

PR 曲線每一個點對應到 Precision（精確率)和 Recall（召回率)。 PR 曲線越平滑，代表模型越好 PR 曲線越接近 (1, 1) 也越好。下圖藍線是使用 CART 模型，綠線使用 SVM 模型，黃線使用 KNN 模型，可以從 PR 曲線看出 CART 模型效能最好。



$$\text{Precision} = \frac{tp}{tp + fp}$$

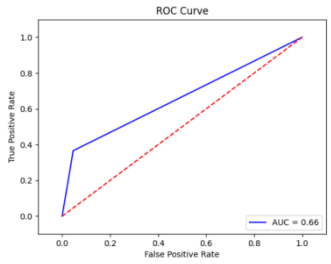
$$\text{Recall} = \frac{tp}{tp + fn}$$

F1 Score

利用 Precision 和 Recall 的調和值（2 * Precision * Recall / (Precision+Recall))來預測 Precision 和 Recall 之間的平衡點，Precision 和 Recall 最接近時，F1 會最大，F1、Precision、Recall 這三者皆越大越好，可以從 PR 曲線看出每個模型使用 F1 的大約數值: CART 0.48、SVM 0.47、KNN 0.46

ROC 曲線

ROC 曲線用圖形來描述二分類模型的效能表現，是一個全面評估模型的指標，無視樣本不平衡的問題(使用 TPR、FPR)。ROC 曲線離對角線越近，模型的準確率越低。



AUC

AUC 是 ROC 曲線裡面的一個判斷指標，代表上圖虛線和藍線的面積，AUC = 1, 是完美分類器，絕大多數的預測場合，不存在完美分類器 0.5 < AUC < 1, 優於隨機猜測，這個模型妥善設定臨界值的話，能有預測價值。 AUC = 0.5 跟隨機猜測一樣（丟銅板），模型沒有預測價值 AUC < 0.5 比隨機預測還差，但只要反預測而行，就優於隨機猜測，這裡的 AUC 為 0.66 。

(五)結論

本報告整理了 3 種監督式學習的模型，並比較其性能和結果，裡面包含 4 種模型指標。傳統方法大多都是針對金流、資產、信用進行分析，使用決策樹，不僅可以在短時間對大數據進行分析，也可以分別對不同種資料進行數據分類，處理的範圍廣，且可以明確看到執行過程，有助於模型的觀察和理解。使用機器學習的方式判斷一個人是否適合繼續借款，可以使銀行可以更快速的對不同的人，不同的資訊做出有效的分析。

(六)參考資料

1. UCI Machine Learning Repository: default of credit card clients Data Set
2. 各常用機器學習(分類)演算法的優缺點總結:DT/ANN/KNN/SVM/GA/Bayes/Adaboosting/Rocchio – jashliao 部落格
3. KNN 演算法優缺點 - IT 閱讀 (itread01.com)
4. 信用分析模型 - MBA 智库百科 (mbalib.com)
5. R 筆記 – (14)Support Vector Machine/Regression(支持向量機 SVM) (rstudio-pubs-static.s3.amazonaws.com)
6. [Day 11] 核模型 - 支持向量機 (SVM) - iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天 (ithome.com.tw)
7. CFA II-Reading 37 信用分析模型. 這篇 Reading 放在 Level... | by Wesley Tzeng | CFA Level 2 Notes | Medium
8. 深度學習分類任務常用評估指標_華為雲開發者社群 - MdEditor (gushiciku.cn)
9. 表現的評估 — 新手村逃脫！初心者的 Python 機器學習攻略 1.0.0 documentation (yaojenkuo.io)
10. 【AI60 問】Q25 決策樹什麼時候使用？有什麼優缺點？ | 緯育 TibaMe Blog
11. KNN 演算法優缺點 - IT 閱讀 (itread01.com)