國立雲林科技大學

資訊管理系碩士班

資料探勘

預測比較 Adult 與 Default of Credit Card Clients 資料集

M11123004 吳浚瑒

M11123017 洪國書

M11123021 李冠穎

M11123028 林永沂

指 導 教 師 ：許 中 川

中 華 民 國 111 年 11 月

摘要

隨著信用卡普及化，各個發卡銀行為了提升市占率，不斷增加發卡數量並降低申請門檻，雖然使得持有信用卡的人很多，但並非所有人都能夠按時還款或全額還款，而還款逾期會導致罰息以及違約金的增加，導致利息越滾越多，陷入債務深淵，這同時造成持卡者以及銀行方的損失。本研究的目的是針對台灣信用卡客戶違約付款案例，數據時間從 2005 年 4 月到 2005 年 9 月，其中包含性別、教育、婚姻狀況、年齡等人口統計數以及賬單金額、上次付款金額、過去付款金額等行為數據，透過對此數據分析來預測違約客戶，本研究採用 Logistic regression、KNN、RandomForest 三種演算法來建構數值預測，結果顯示 Logistic regression 有大約 87%的正確率，KNN 有大約 90%的正確率，Random Forest 有大約 88%的正確率，綜合結果發現 KNN 預測的結果與績效具有高度參考價值，能較為精準預測違約客戶，此外本研究還使用了成人(Adult)數據集，其中包含年齡、職業別、學歷…等欄位資料，目的是以這些欄位預測每週的工作小時數。

Keyword：信用卡、還款逾期、違約客戶、Adult 數據集、Default

壹、緒論

1.1 研究背景

信用卡(Credit Card)是一種非現金付款的交易方式，在消費時無須支付現金，只需在賬單日時再進行還款。起源自十九世紀末英國，依賴富裕人口的資本信用而設計，針對富裕人口購買昂貴奢侈品時但現金不足的情形。當時信用卡概念僅能於特定場所進行短期的商業交易行為，無法長期積欠欠款，並且須盡速還清。

第一張針對大眾的信用卡出現在 20 世紀 50 年代。在 1949 年，美國曼哈頓信貸專家 Frank McNamara 於飯店用餐時，因現金不足，只能請家人幫忙帶現金來支付的情況感到困擾而興起的念頭，在次年便創立了大來俱樂部(Diner's Club)，在初期，大來俱樂部主要與餐廳合作，會員可持俱樂部提供之記帳卡於紐約的 27 家合作餐廳簽約消費，款項先由俱樂部墊付給店家，每月再與持卡會員結帳，後來合作逐漸擴展至飯店、航空公司及旅遊業等各個行業。這種先消費，後付款的商業模式在後來的日常生活中廣泛運用，建立了現代信用卡的概念雛形。

而在台灣，第一張信用卡為民營中國信託公司在民國 63 年所發行的「信託信用卡」，但此時並不具備循環信用功能，只能算是「簽帳卡」。直到民國 72 年，聯合信用卡處理中心成立，並於次年發行「聯合簽帳卡」，與此同時，結束原先發行「信託信用卡」之服務。民國 78 年 VISA 國際組織進入台灣，並與聯合信用卡處理中心合作推出「VISA 國際信用卡」，馬上就吸引國內信用卡發卡銀行與國外信用卡發卡銀行展開國際競爭，民國 82 年財政部放寬對信用卡的管制，開放「循環信用」功能，此時我國信用卡才真正從簽帳卡型態轉型成真正具借貸功能的信用卡。

時至今日，據金管會銀行局於 2022 年 9 月的信用卡業務統計可以看出，本國銀行信用卡之流通卡數高達 5 千 4 百多萬張(金管會銀行局，2022)，證明了信用卡已普及使用於民眾日常生活中。

隨著信用卡普及化，各個發卡銀行為了提升市占率，不斷增加發卡數量並降低申請門檻，雖然使得持有信用卡的人很多，但並非所有人都能夠按時還款或全額還款，而還款逾期會導致罰息以及違約金的增加，導致利息越滾越多，陷入債務深淵，這同時造成持卡者以及銀行方的損失。為了降低銀行風險，本研究資料蒐集從 UCIDataSets 找到國內信用卡的違約資料集，透過資料探勘技術根據性別、年齡、婚姻狀況、過去交易等數據來預測違約客戶。

Adult 資料集是資料探勘的資料集，資料是從美國 1994 年人口普查資料庫中抽取而來，資料集已經劃分為訓練資料。其中變數包括年齡、學歷、職業等重要資訊，其中有屬於類別離散型變數，另外也有屬於數值連續型變數。該資料集是一個分類資料集，用來預測年收入(Adult 資料集分析及四種模型實現，2020)。

1.2 研究目的

本研究的目的是針對台灣信用卡客戶違約付款案例，從 UCIDataSets 的信用卡違約資料集蒐集資料，並利用 python 來分析該數據，數據時間從 2005 年 4 月到 2005 年 9 月，其中包含性別、教育、婚姻狀況、年齡等人口統計數，以及帳單金額、上次付款金額、過去付款金額等行為數據，透過對此數據分析來預測違約客戶，並比較不同學習方法之差異。

Adult 資料集的目的是預測每周工作時間，從 Adult 資料集中利用 Python 來分析資料，數據時間從美國 1994 年人口普查資料庫中抽取而來，其中包含變數包括年齡、學歷、職業等重要數據，透過此數據來預測每周工作時間。

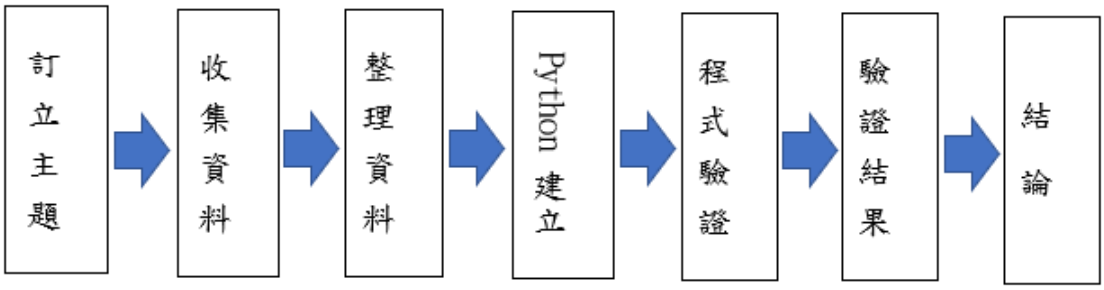
貳、方法

2.1 研究架構

本研究的研究架構(**圖 1**) 如下，將數據集的資料依序步驟進行收集、整理、程式建置、驗證、結論。試著將信用卡數據集內的資料利用 Python 製作三種不同的演算法並建構預測模型，得到的結論將有助於日後研究有關預測信用卡客戶的數據參考。

圖 1

研究架構



2.2 執行程式的方法

(1) 選擇資料集

從 UCI Dataset 中挑選資料集，並將挑選出來的變數簡化為目標資料集。

(2) 預處理

針對資料的完整性去過濾不符合規則的資料，例如”空值”。

(3) 變換

將資料的輸入格式轉換成符合後續資料探勘步驟的格式。

(4) 資料探勘

資料經過轉換後，使用者可以藉著事先預定好的步驟，並藉由一種或多種的技術，來萃取出資料中隱含的重要訊息及模式。

(5) 解釋/評估

資料經由萃取，並經由統計或其他技術來確認其結果，且所萃取資訊描述範圍可以擴展到資料庫中未曾察覺或包含的資料。

參、實驗

3.1 資料集

本實驗所選擇的資料分別是 Adult DataSet、Default of credit card clients DataSet

1. AdultDataSet 有 48842 筆資料、15 個欄位，分別是 Age、Workclass、fnlwgt、Education、Education-num、Marital_Status、Occupation、Relationship、Race、Sex、Capital-gain、Capital-loss、hrs_per_week、Native-Country、Earning_potential。如圖 2。

2. Default of credit card clients Dataset 有 30000 筆資料、25 個欄位，分別是 Amount of the given credit 、Gender、Education、Marital status 、Age、PAY_0: Repayment status in September, 2005、PAY_2: Repayment status in August, 2005、PAY_3: Repayment status in July, 2005、PAY_4: Repayment status in June, 2005、PAY_5: Repayment status in May, 2005、PAY_6: Repayment status in April, 2005、BILL_AMT1: Amount of bill statement in September, 2005、BILL_AMT2: Amount of bill statement in August, 2005、BILL_AMT3: Amount of bill statement in July, 2005、BILL_AMT4: Amount of bill statement in June, 2005、BILL_AMT5: Amount of bill statement in May, 2005、BILL_AMT6: Amount of bill statement in April, 2005、PAY_AMT1: Amount of previous payment in September, 2005、PAY_AMT2: Amount of previous payment in August, 2005、PAY_AMT3: Amount of previous payment in July, 2005、PAY_AMT4: Amount of previous payment in June, 2005、PAY_AMT5: Amount of previous payment in May, 2005、PAY_AMT6: Amount of previous payment in April, 2005。如圖 3。

	Age	Workclass	fnlwgt	Education	Education-num	Marital_Status	Occupation	Relationship	Race	Sex	Capital-gain	Capital-loss	hrs_per_week	Native-Country	Earning_potential
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K.
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K.
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K.
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K.
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K.

圖 2　AdultDataSet

index	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2
0	1	20000	1	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689
1	2	120000	1	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000
2	3	90000	1	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500
3	4	50000	1	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019
4	5	50000	0	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681

圖 3　Default of credit card clients DataSet

3.2 實驗設計

Default of credit card dataset

首先 Import 本實驗會使用到的 Python 函式，把 Excel 裡面的欄位名稱輸入進去並且列出表格，接著移除了 ID 不重要的欄位，並且做 Data Clening。在跑演算法之前，本實驗藉由 Feature Engineering 提高準確率、優化收斂速度，演算法本實驗將利用 Logistic regression、KNN、RandomForest 來建構數值預測，然後再用 RMSE（Root Mean Square Error）、MAE（Mean Absolute Error）來比較各個演算法預測績效。

Adult Data Set

首先採用 pandas 和 Numpy 來進行預處理，來讀取資料是否有缺失值(空值)，將缺失值轉換成 nan 跟 NaT，本實驗，本實驗使用三種模型進行分析，分別是 Decision Tree Classifier、Support Vector Classifier、k-nearest neighbors algorithm (k-NN)。

3.3 實驗結果

Default of credit card dataset

透過以上實驗方式，進行異常值處理，同時糾正數據不平衡，並使用 SMOTE 特徵選擇，Logistic regression 提供大約 87%的正確率，KNN 提供大約 90%的正確率，Random Forest 提供大約 88%的正確率，本研究使用 RMSE 作為績效指標，分別是 0.30154、0.31464、0.33182，綜合評估下 KNN 預測出來的正確率以及績效是值得參考的。

	Train Accuracy	Test Accuracy	F1 Score	Recall Score	Precision Score	Roc Auc Score
Logistic Regression Model	0.8684	0.8798	0.6815	0.8329	0.5767	0.9468
K Nearest Neighbor Model	0.9115	0.9010	0.6542	0.8646	0.533	0.9192
Random Forest Model	0.9997	0.8706	0.6269	0.7976	0.5884	0.9475

表 1　三種演算法比對表格

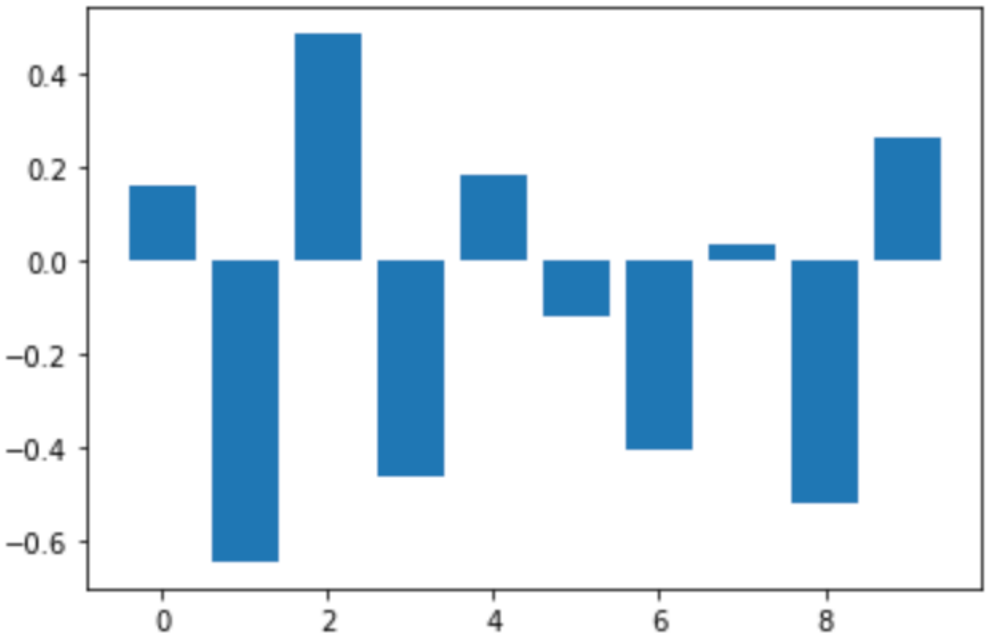


圖 4　Feature importance

Adult DataSet

透過以上實驗方式，KNN、SVC、Decision Tree 所得出的正確率分別是 83%、85%、76%，並且透過 RMSE 去評估各個演算法的績效，分別是 0.4104、0.3865、0.3865，綜合以上數據，SVC 預測的數值與績效是值得參考的。

根據圖 5 以及圖 6 顯示，很多人每週平均工時大約 40 小時，範圍從 30-50 小時不等。

	Train Score	**Test Score**
Support Vector Classifier	**0.840832**	**0.833282**
K Nearest Neighbor Model	**0.842982**	**0.837581**
Decision Tree Model	**0.763130**	**0.766349**

表 2　AdultDataSet 訓練測試結果

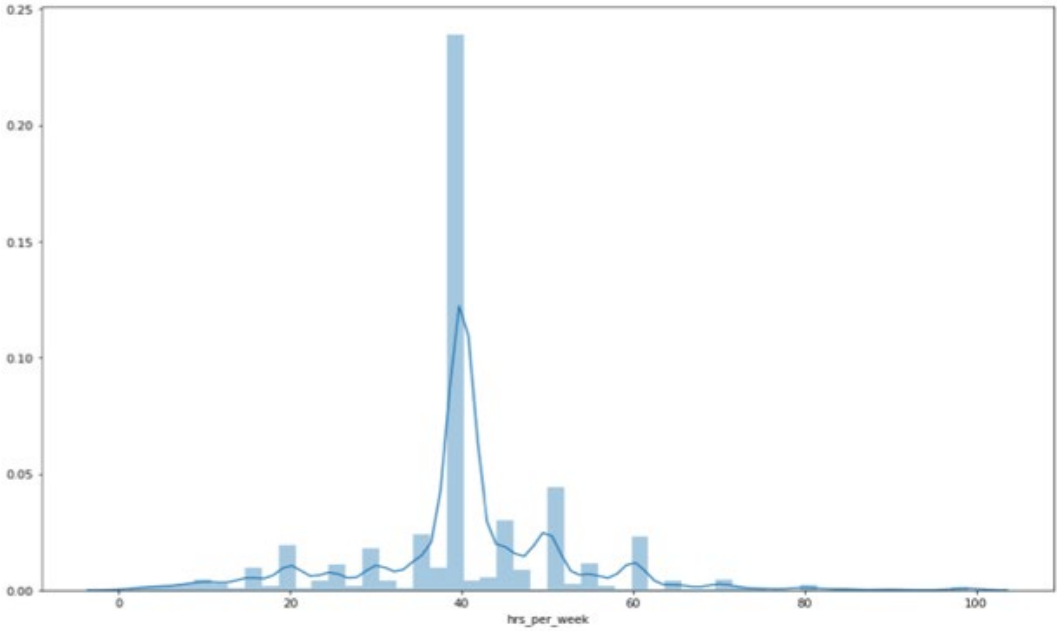


圖 5　平均工時統計

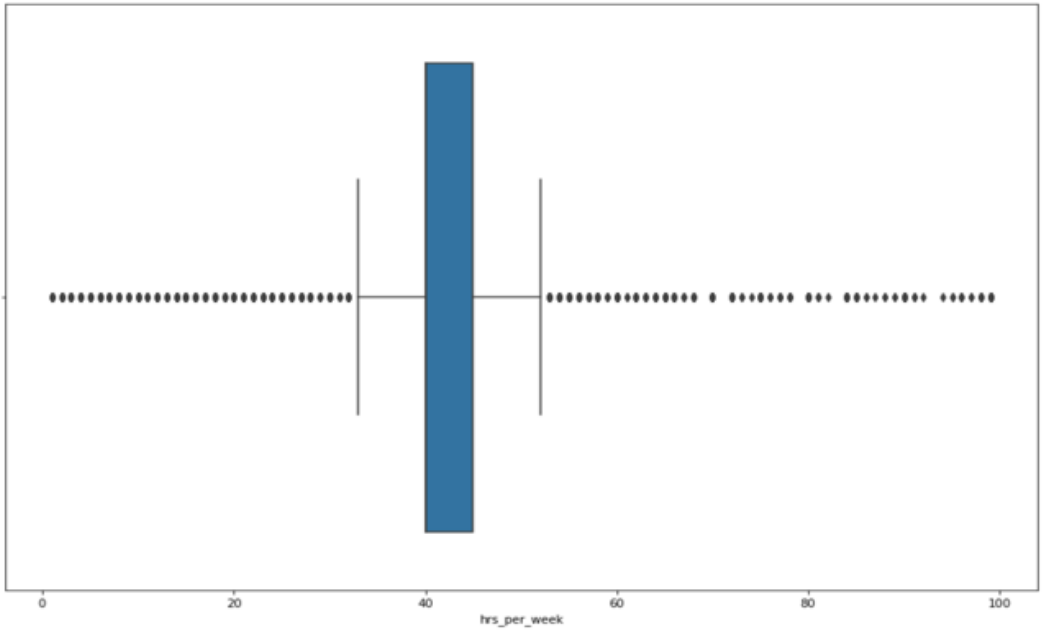


圖 6　平均工時區間

肆、結論

本研究對 Default of credit card dataset 進行不同的演算法進行了檢查並展示了它們的細節，並構建一個分類模型，得出各模型的正確率以及績效評估並且也計算了各特徵的重要性，該模型能夠預測信用卡客戶是否會在下個月違約。Adult dataset 的部分，本研究也進行三種不同的演算法來評估各個模型預測數值的差異，針對 hours per week 欄位的部分，本研究發現人們傾向於每週工作 40 小時。

參考文獻

金融監督管理委員會 銀行局(金管會銀行局) 2022 年 9 月

https://www.banking.gov.tw/ch/home.jsp?id=157&parentpath=0,4&mcustomize=bstatistics_view.jsp&serno=201105120008

Adult 資料集分析及四種模型實現. (2020, March 3). Gushiciku.Cn.

https://www.gushiciku.cn/pl/aV5F/zh-tw

Default of credit card clients Data Set

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients