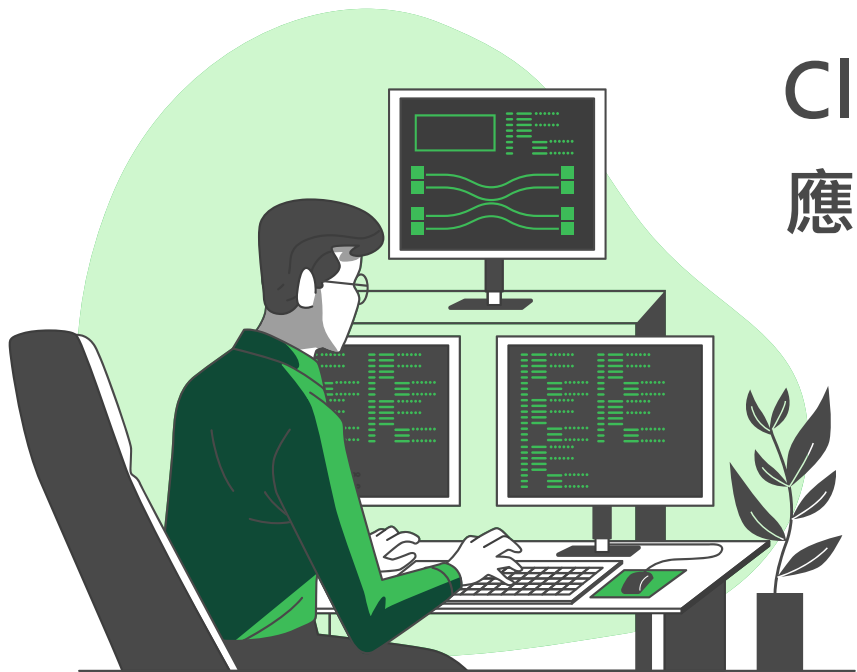


Cluster-than-Predict方法 應用於客戶信貸違約預測

陳怡仁 2023.08.16



摘要

根據金管會於2024.06.20發佈之「金融業運用人工智慧（AI）指引」，建議金融機構對風險程度較高之AI系統建立可解釋性原則，而XGBoost作為一款強大的分類模型，其黑箱性質違反此原則。相比之下，本研究的先分群後預測方法更為簡單易懂，經過資料前處理與特徵交乘後的AUC、Precision甚至超越相同狀態下的XGBoost。

除此之外，本研究使用Tableau軟體進行資料視覺化分析，在帶入模型後也輸出PCA、群內違約率、Entropy等比較圖，在預測客戶是否違約時，也能夠進行客戶分群，為後續的利率訂定、客製化商品、風險分層、客戶管理奠定基礎，以輔助商業決策。

目錄

01

...

Data

資料來源
資料視覺化
資料分析

02

...

Method

資料前處理
先分群後預測
程式流程圖

03

...

Result

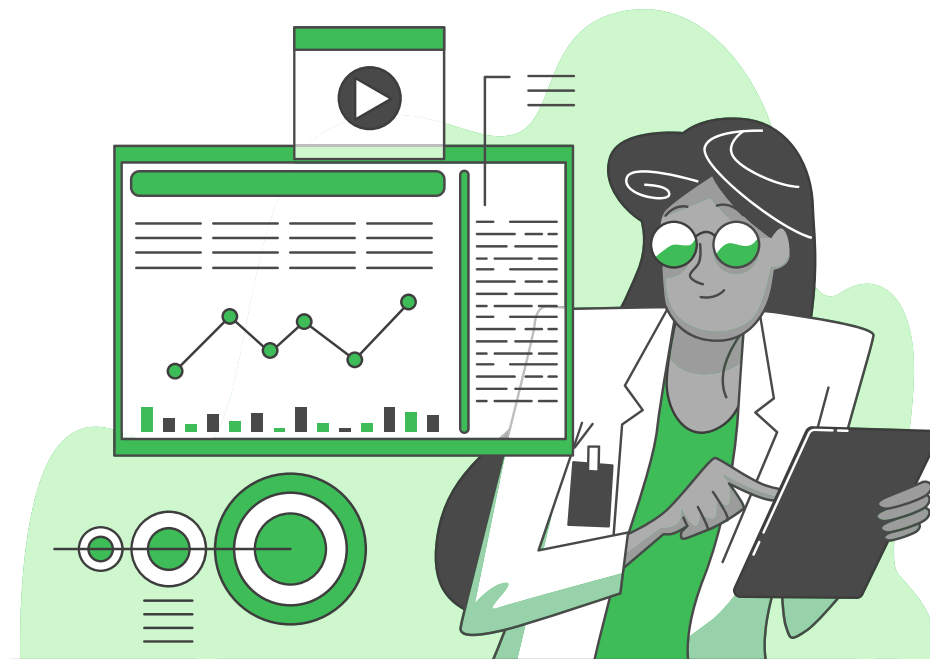
分群結果分析
預測結果比較

04

...

Summary

結果與討論
未來工作



01 Data

資料來源
資料視覺化
資料分析

資料來源

- 資料名稱：Binary Classification with a Bank Churn Dataset
- 時間：2024.01.02~2024.01.31
- SIZE：13欄× 165034列
- 特徵：id、CustomerId、Surname、CreditScore、Geography、Gender、Age、Tenure、Balance、NumOfProducts、HasCrCard、IsActiveMember、EstimatedSalary、Exited
- 網址：<https://www.kaggle.com/competitions/playground-series-s4e1>

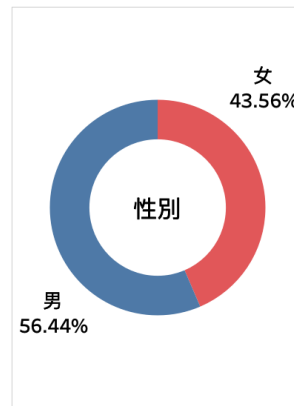
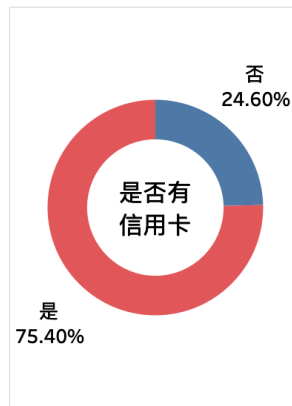
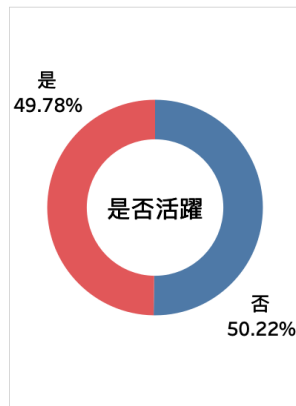
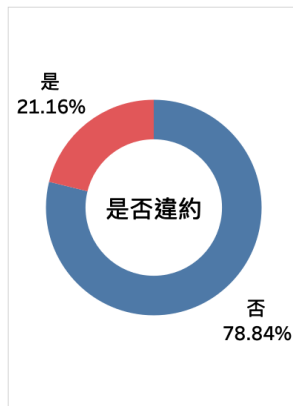


Binary Classification with a Bank Churn Dataset

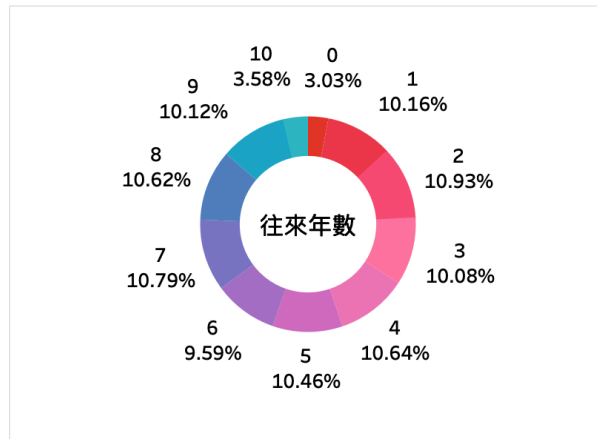
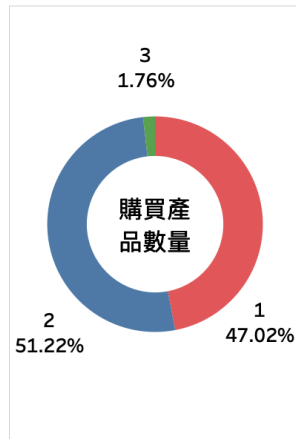
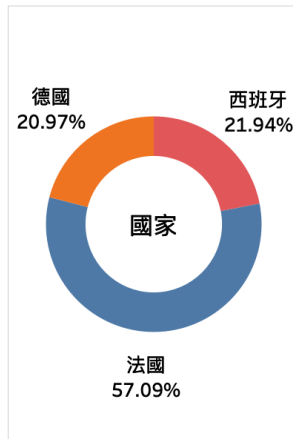
Playground Series - Season 4, Episode 1

Playground · 3632 Teams · 6 months ago

資料視覺化-離散型資料



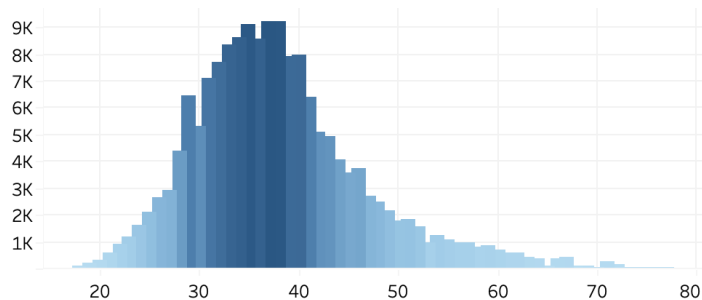
是否違約
 否
 是
 是否活躍
 否
 是
 是否有信用卡
 否
 是
 性別
 女
 男



國家
 西班牙
 法國
 德國
 購買產品數量
 1
 2
 3
 往來年數
 0
 1
 2
 3
 4
 5
 6
 7
 8
 9
 10

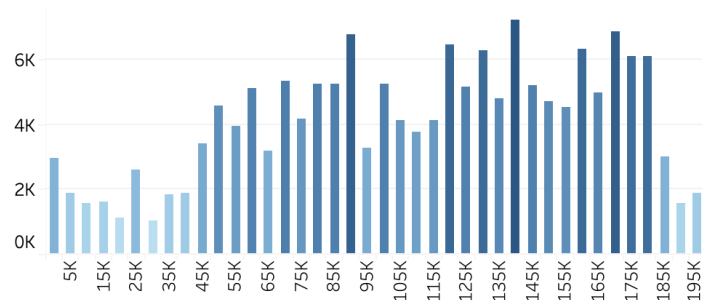
資料視覺化-連續型資料

年齡



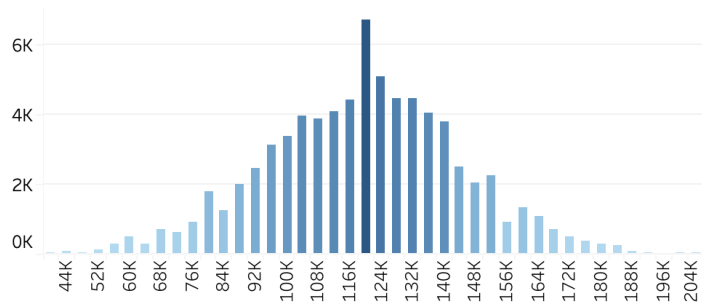
計數 33 9,255

薪資



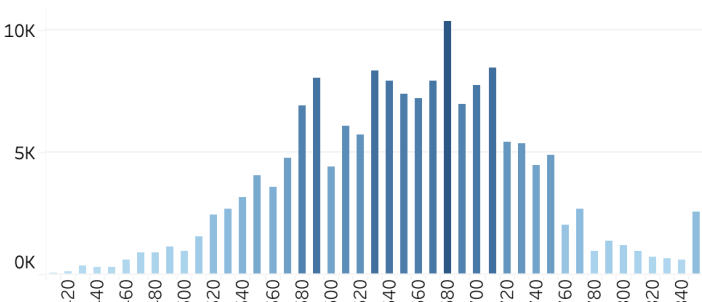
計數 1,032 7,204

帳戶餘額



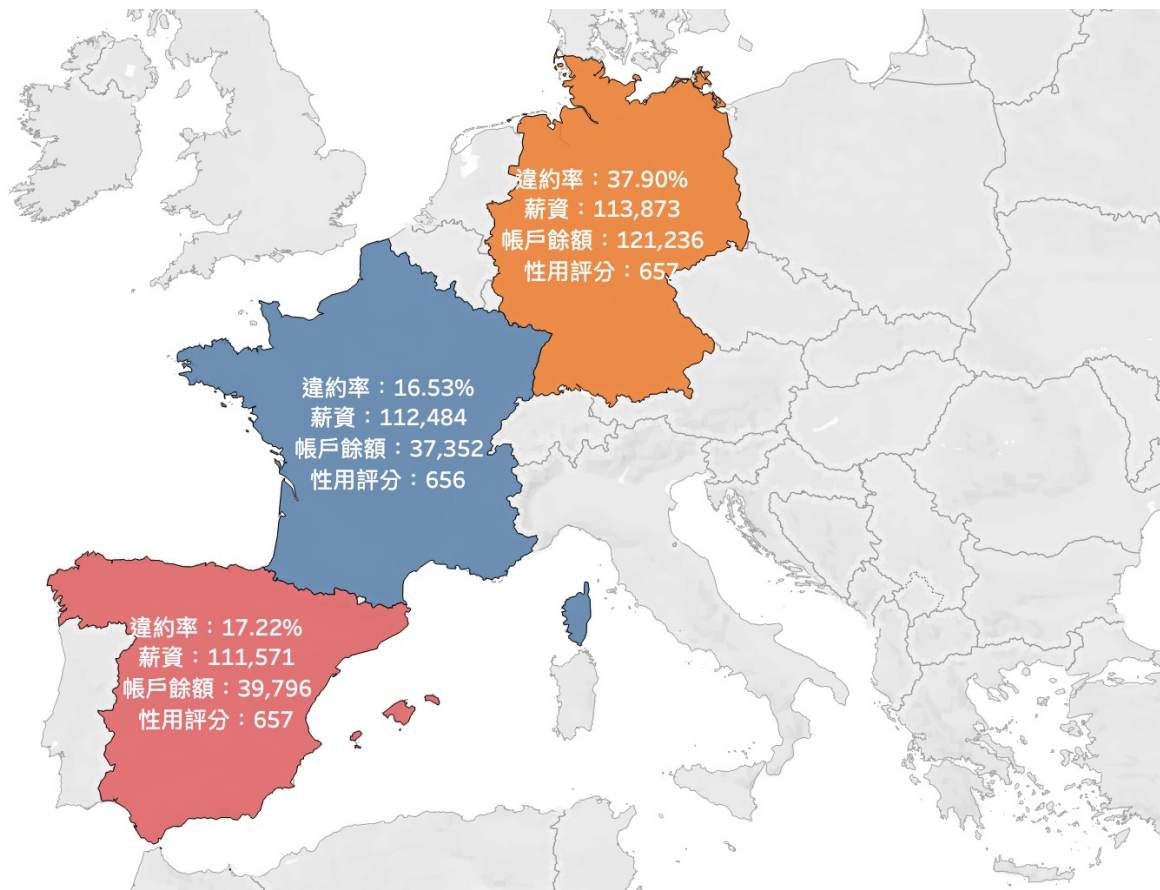
計數 33 6,725

信用評分



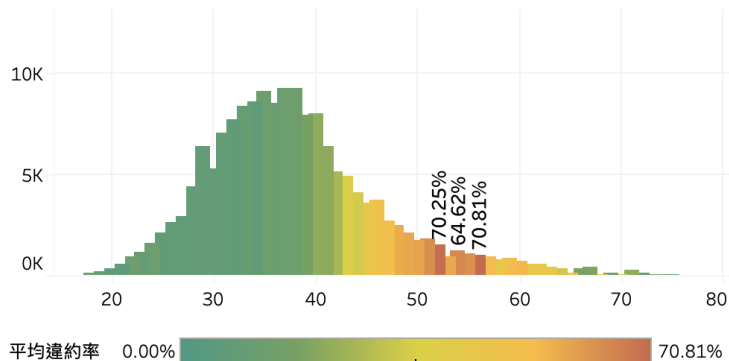
計數 84 10,377

資料分析-國家平均數據

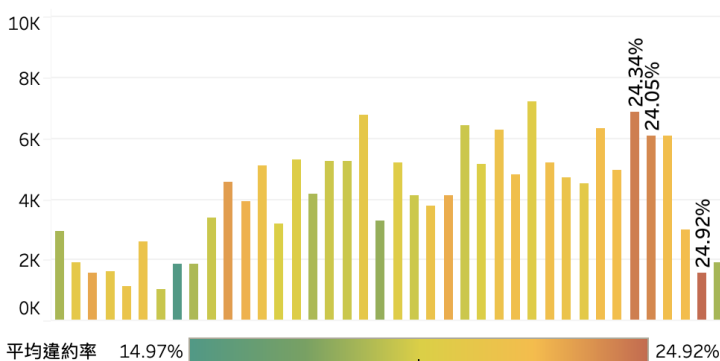


資料分析-連續型資料的違約分佈

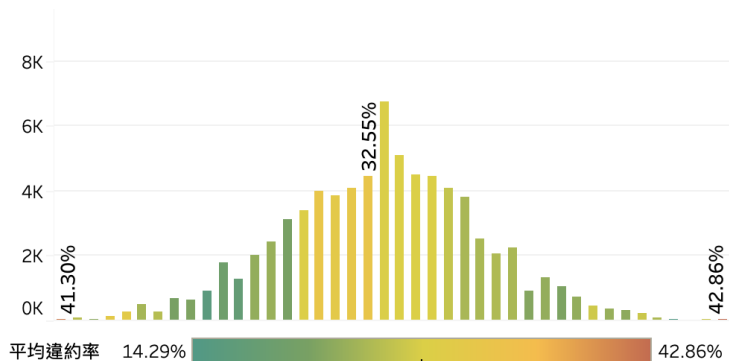
年齡分析



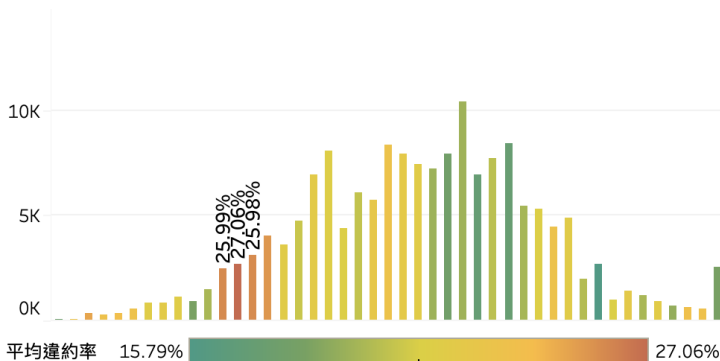
薪資分析



帳戶餘額分析



信用評分分析




資料分析-離散型資料的違約分佈

購買產品數量分析

年齡(群組)	女				男			
	1	2	3	4	1	2	3	4
18~27	19.08%	4.44%	77.97%		11.37%	2.47%	65.96%	
28~36	21.95%	3.87%	84.42%	85.29%	12.86%	2.01%	72.83%	76.32%
37~45	44.88%	9.19%	92.92%	89.19%	27.65%	4.79%	83.87%	79.27%
46~55	76.37%	28.40%	94.29%	95.40%	61.99%	18.66%	93.55%	90.16%
56~65	71.69%	24.59%	95.12%	94.12%	55.79%	16.13%	93.70%	
66~75	41.88%	9.09%			22.01%	4.77%		
76~92					18.18%	0.00%		

高風險！

購買四個產品的46~55歲女性
購買三個產品的56~65歲男性

平均違約率 0.00%  95.40%

是否活躍分析

年齡(群組)	女		男	
	否	是	否	是
18~27	15.44%	7.29%	8.98%	4.11%
28~36	17.46%	7.48%	10.17%	4.09%
37~45	38.58%	17.57%	23.19%	9.37%
46~55	77.12%	45.41%	65.27%	30.19%
56~65	86.62%	37.85%	76.56%	25.24%
66~75	67.70%	16.69%	44.44%	8.50%
76~92		2.50%		4.71%

是否有信用卡分析



年齡(群組)	女		男	
	否	是	否	是
18~27	12.54%	10.91%	9.00%	5.64%
28~36	14.21%	11.79%	8.37%	6.54%
37~45	30.65%	28.45%	18.69%	15.74%
46~55	61.73%	65.80%	47.47%	49.81%
56~65	58.08%	59.53%	44.70%	42.02%
66~75	26.46%	27.96%	16.82%	12.73%
76~92		2.86%	12.90%	4.41%

高風險！

非活躍客戶的56~65女性

高風險！

有信用卡的46~55歲女性

平均違約率 2.50%  86.62% 平均違約率 2.86%  65.80%

資料分析-離散型資料的違約分佈

往來年數分析

年齡(群組)	女										
	0	1	2	3	4	5	6	7	8	9	10
18~27	11.90%	14.84%	10.02%	11.72%	11.54%	10.64%	12.28%	10.07%	9.88%	11.57%	11.17%
28~36	15.98%	12.70%	11.13%	13.69%	12.86%	13.86%	10.76%	10.60%	11.68%	12.42%	14.67%
37~45	36.13%	30.03%	27.40%	32.37%	30.72%	28.34%	26.77%	25.88%	28.41%	28.68%	30.57%
46~55	68.27%	65.43%	61.40%	66.51%	66.84%	66.57%	63.86%	62.11%	62.22%	66.91%	63.49%
56~65	53.49%	59.02%	58.21%	62.22%	59.18%	62.95%	57.09%	55.31%	57.58%	62.69%	57.76%
66~75		27.50%	20.78%	27.17%	28.77%	30.00%	21.95%	34.57%	23.53%	30.67%	26.47%

高風險！

往來小於一年的
46~55歲女性

往來年數分析

年齡(群組)	男										
	0	1	2	3	4	5	6	7	8	9	10
18~27	9.09%	6.05%	5.00%	6.79%	7.88%	7.89%	6.47%	5.77%	5.53%	6.50%	5.81%
28~36	9.87%	7.15%	6.39%	7.44%	7.70%	7.05%	6.36%	6.28%	6.03%	7.88%	6.46%
37~45	20.66%	17.74%	14.94%	17.27%	18.36%	16.99%	14.97%	15.10%	15.64%	15.77%	16.96%
46~55	50.17%	49.85%	47.59%	53.29%	50.79%	51.77%	48.96%	46.21%	46.11%	49.23%	43.66%
56~65	39.81%	46.90%	38.99%	45.21%	45.83%	38.16%	49.31%	41.54%	43.23%	38.18%	41.49%
66~75		13.98%	16.19%	10.87%	15.24%	12.26%	17.65%	9.09%	14.61%	13.79%	10.00%

高風險！

往來三年的
46~55歲男性

02 Method

資料前處理
先分群後預測
程式流程圖

資料前處理

- One Hot Encoding

ID	性別
1	女
2	男
3	女



ID	男	女
1	0	1
2	1	0
3	0	1

- Z標準化, $Z = \frac{X - \mu}{\sigma}$

X_1	X_2
3.3	0
1.6	3
2.4	2



X_1	X_2
1.02	-1.09
-0.98	0.87
-0.39	0.22

- 特徵交乘
(以二次交乘為例)

X_1	X_2
3.3	0
1.6	3
2.4	2

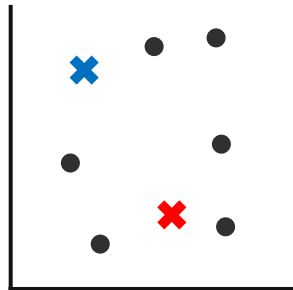


X_1	X_2	X_1^2	X_2^2	X_1X_2
3.3	0	10.89	0	3.3
1.6	3	2.56	9	4.8
2.4	2	5.76	4	4.8

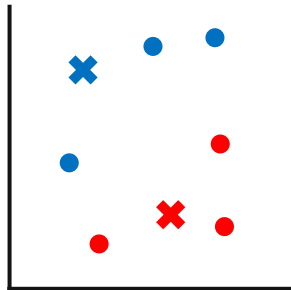
先分群後預測

- K-means分群

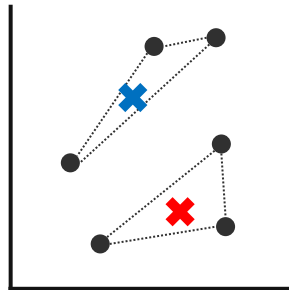
隨機選2個中心點



按距離分組



計算質心當作新中心點

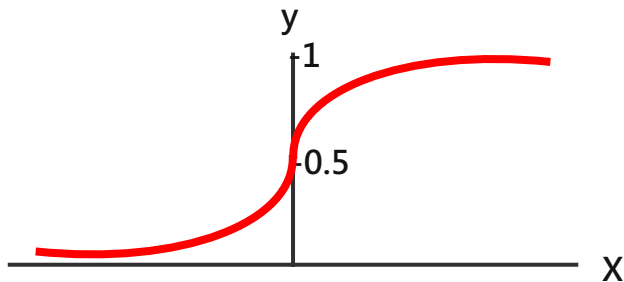


重複執行直到中心點不變

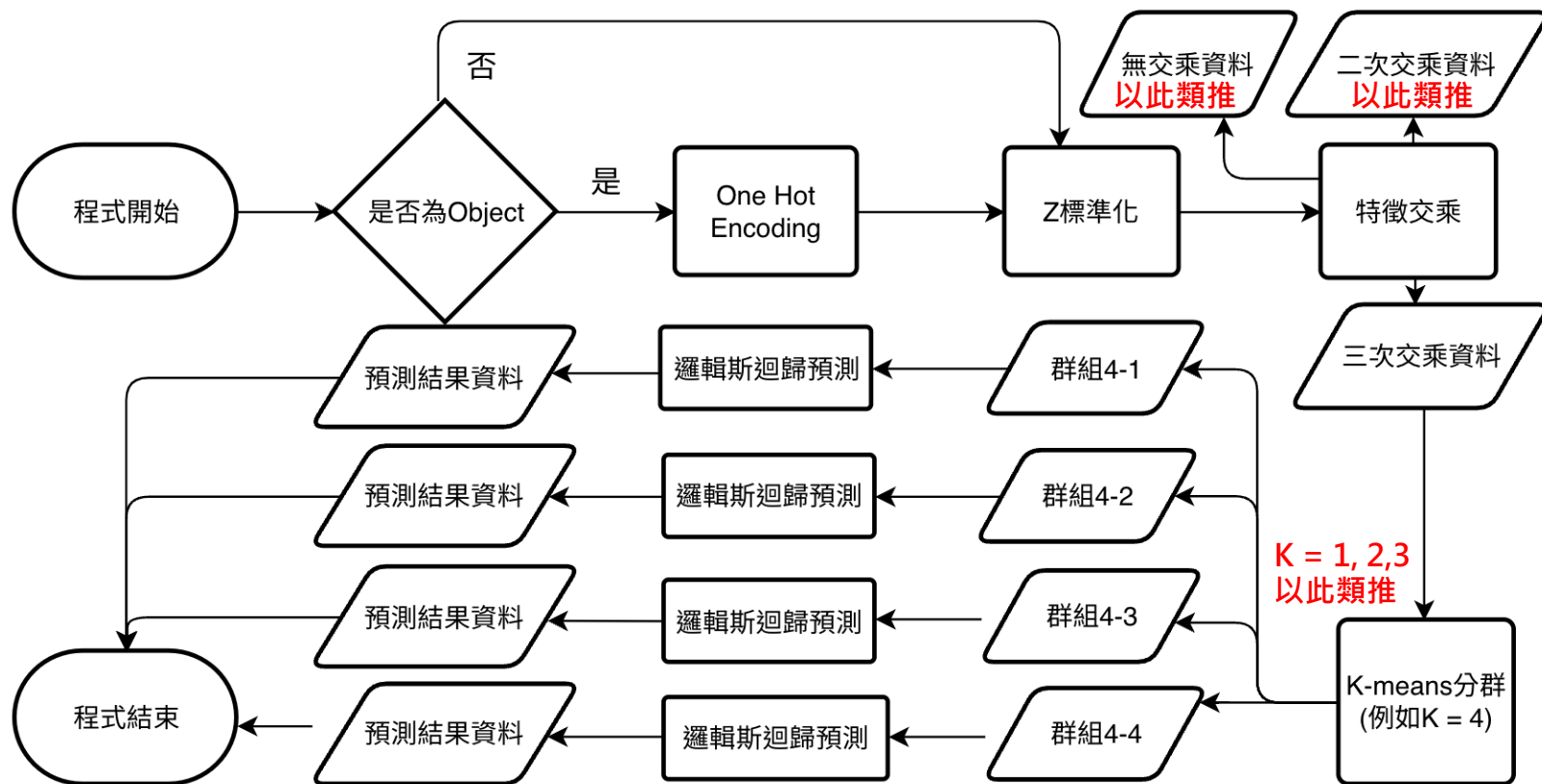
- 邏輯斯迴歸

$$y = \frac{1}{1 + e^{(\beta_0 + \beta_1 X)}}$$

適用於二元分類問題



程式流程圖（以三次交乘分四群為例）



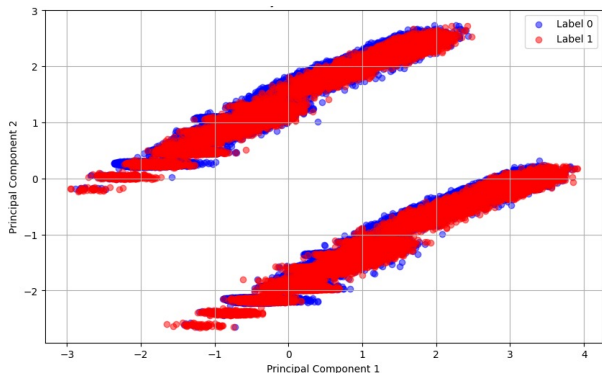
03 Result

分群結果分析
預測結果比較

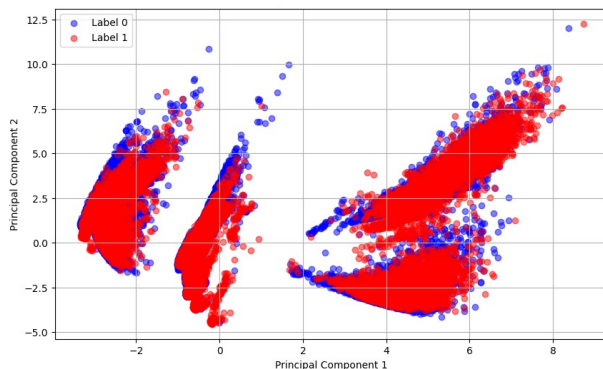
分群結果分析-PCA

- 藍點代表負案例，紅點代表正案例，透過主成分分析將特徵濃縮成二維平面
- 無論是無交乘、二次交乘、三次交乘皆有明顯的群集現象，適用分群方法
- 正負案例皆均勻分散至各群，並無因為數據不平衡引發的正案例離群現象

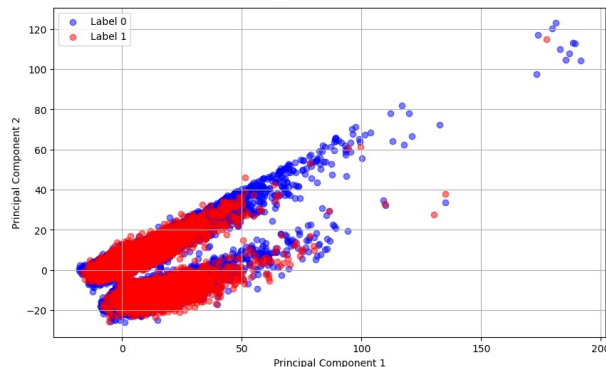
無交乘



二次交乘



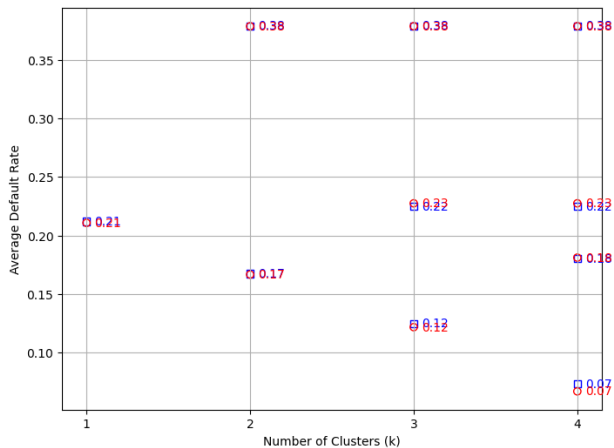
三次交乘



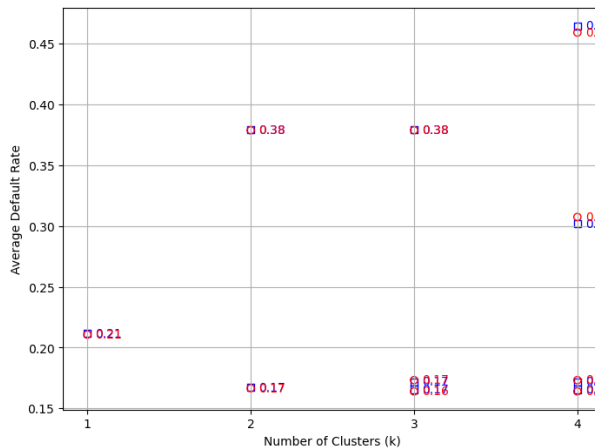
分群結果分析-群內違約率（越分散越好）

- 藍字代表訓練集，紅字代表測試集，無交乘的訓練/測試集群內違約率皆相似
- 無交乘和三次交乘的群內違約率較分散，分群結果較有區別性，二次交乘則無

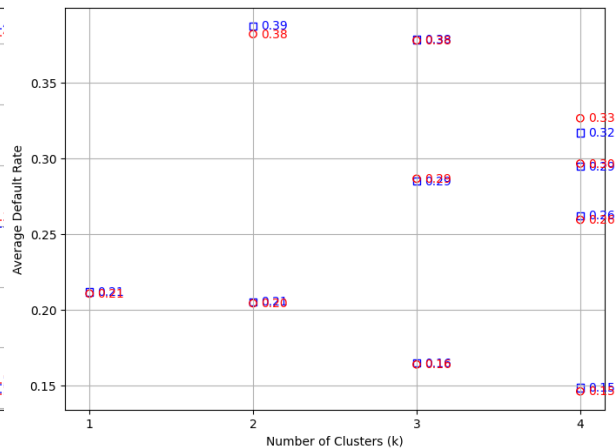
無交乘



二次交乘



三次交乘



分群結果分析-Entropy (越低越好)

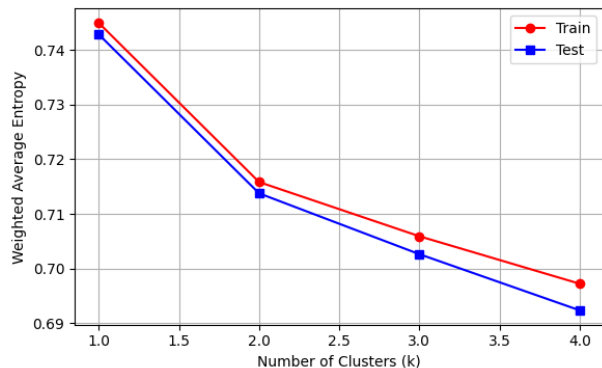
- 無交成和二次交乘的Entropy與分群數量呈負相關，代表訊息複雜度越低，群內資料更具同值性

$$H(C_i) = \sum_j P_{ij} \times \log_2(P_{ij})$$
$$H_{weighted} = \sum_{i=1}^k \frac{N_i}{N} \times H(C_i)$$

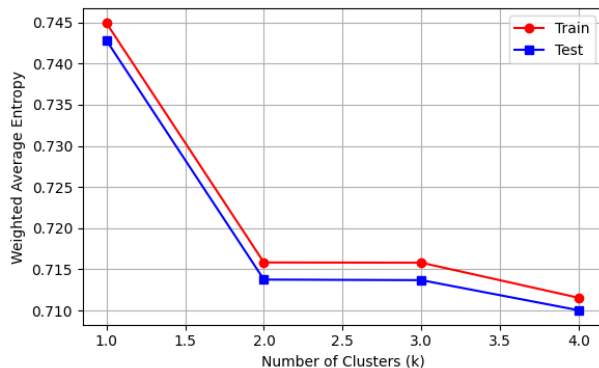
$\frac{N_i}{N}$ 是群 C_i 在整個數據集中的樣本比例

$H(C_i)$ 是群 C_i 的 *Entropy*

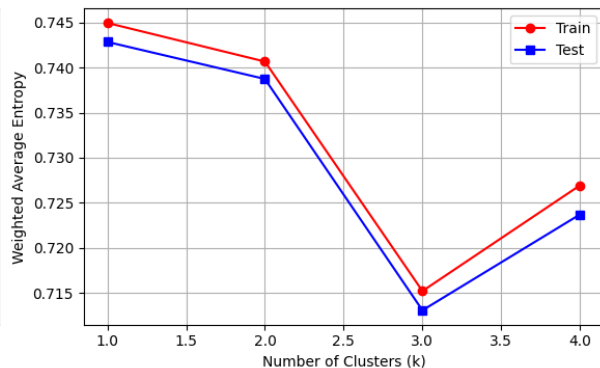
無交乘



二次交乘



三次交乘



預測分析-AUC (越高越好)

- LR在二次和三次交乘下的AUC勝過XGB
- LR的AUC與交乘數呈正相關，XGB呈負相關
- LR的AUC與分群數無相關，XGB呈負相關

$$AUC = \frac{\sum \Psi(P_{\text{正樣本}}, P_{\text{負樣本}})}{m \times n} \quad \Psi = \begin{cases} 1, & P_{\text{正樣本}} > P_{\text{負樣本}} \\ 0.5, & P_{\text{正樣本}} = P_{\text{負樣本}} \\ -1, & P_{\text{正樣本}} < P_{\text{負樣本}} \end{cases}$$

$m \times n$ 為正樣本數 \times 負樣本數

模型	無交乘		二次交乘		三次交乘	
	LR	XGB	LR	XGB	LR	XGB
K = 1	0.8180	0.8886	0.8814	0.8852	0.8886	0.8837
K = 2	0.8208	0.8862	0.8824	0.8814	0.8878	0.8778
K = 3	0.8212	0.8832	0.8823	0.8778	0.8875	0.8753
K = 4	0.8230	0.8812	0.8822	0.8754	0.8863	0.8705

預測分析-Precision (越高越好)

- LR在二次和三次交乘下的Precision勝過XGB
- LR的Precision與交乘數呈正相關，XGB呈負相關
- LR的Precision與分群數無相關，XGB呈負相關

$$Precision = \frac{\text{將正類預測為正類的數量}}{\text{預測的正類數量}}$$

模型	無交乘		二次交乘		三次交乘	
	LR	XGB	LR	XGB	LR	XGB
K = 1	0.6964	0.7375	0.7393	0.7320	0.7453	0.7245
K = 2	0.7028	0.7273	0.7373	0.7193	0.7409	0.7230
K = 3	0.7021	0.7241	0.7383	0.7134	0.7447	0.7145
K = 4	0.7089	0.7181	0.7396	0.7122	0.7380	0.7049

預測分析-Recall (越高越好)

- LR在所有交乘下的Recall皆無勝過XGB
- LR的Recall與交乘數呈正相關，XGB呈負相關
- LR的Recall與分群數無相關，XGB呈負相關

$$Recall = \frac{\text{將正類預測為正類的數量}}{\text{原本的正類數量}}$$

模型	無交乘		二次交乘		三次交乘	
	LR	XGB	LR	XGB	LR	XGB
K = 1	0.3882	0.5668	0.5346	0.5554	0.5487	0.5577
K = 2	0.3951	0.5615	0.5356	0.5549	0.5490	0.5561
K = 3	0.3955	0.5618	0.5359	0.5528	0.5524	0.5556
K = 4	0.4083	0.5592	0.5334	0.5477	0.5511	0.5525

04 Summary

結果與討論
未來工作

結果與討論

在完整、分布正常、維度小的二元分類問題中，經過特徵交乘的先分群後預測方法，在大部分情況下，能使邏輯斯迴歸的AUC、Precision優於相同條件下的XGBoost。

模型	AUC		Precision		Recall	
	LR	XGB	LR	XGB	LR	XGB
勝負	二次和三次交乘	無交乘	二次和三次交乘	無交乘		所有交乘
交乘數	正相關	負相關	正相關	負相關	正相關	負相關
分群數	無相關	負相關	無相關	負相關	無相關	負相關

未來工作

在**不完整**、**分布不正常**、**維度大**的二元分類問題中，經過特徵交乘的先分群後預測方法，在大部分情況下，能使邏輯斯迴歸的AUC、Precision優於相同條件下的XGBoost。

不完整



填充缺失值

- 迴歸填充法
- KNN填充
- 隨機填充法

分布不正常



離群值處理

- 百分位刪除法
- 對數轉換
- 孤立森林法

維度大



特徵選擇

- 過濾法
- 包裝法
- 嵌入法



Thanks!

陳怡仁 2023.08.16

Mail : pauljkk20001009@gmail.com

https://github.com/Pauljkk/Cluster_than_Predict_by_ChenYiJen