信用卡客戶統計建模

壹、Recape

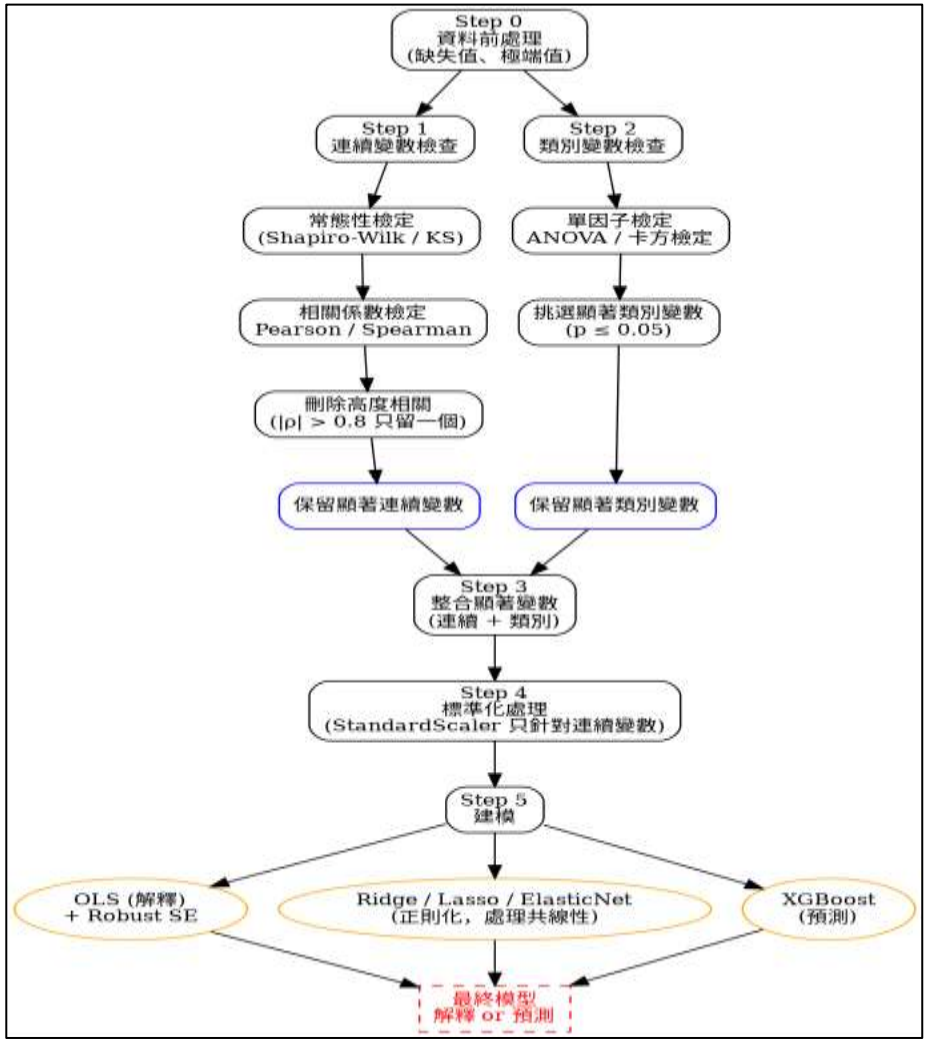
1. 延續七月的結論建議，因上次為連續變數間相關性，如：卡片額度與未出帳帳單金額相關性為 0.55、卡片額度與貢獻度相關性為 0.415 的中度相關，表示**有刷卡行為來的比申辦信用卡**還重要。
2. 歸納以上一年期間信用卡刷卡紀錄，明顯觀察出南北信用卡的業績量差異，除**成大、屏東**兩南部分行，其餘皆為台中以北，主要需與它行競爭外，與分行家數也有正相關，或許初步可針對這兩間南部分行加強行銷力度，如：百貨或者大型超市駐點推廣、當地人個性及喜好。
3. 後續針對資料源報表相關日期欄位釐清，以便正確抓取資料，後續進一步透過統計方式針對類別及連續兩種變數篩選影響刷卡動機的顯著因子，針對其變數做解釋或預測。

貳、商業問題

此次針對上月資料，使用客戶**刷卡筆數**作為反應變數，因刷卡筆數較刷卡金額較趨近於常態分佈，故以該變數作為 y；另，分別針對類別型資料(幣別、消費地點名稱、客戶所屬分行名稱、年齡區間)、連續型資料(前一期繳款金額、未出帳金額、刷卡金額)做資料前處理(dummy, standardize) → 單因子分析(顯查顯著性) → 多因子分析及關係係數分析(檢查共、顯著性) → 正則化迴歸(Lasso,Ridge)等…篩選合適模型(若使用傳統最小平方法或機器學習可省略這段，如：OLS,XGBoost)。

參、建模過程

1. 查找資料源(Tableau) →
2. 資料前處理：連續：標準化、合併小樣本、NaN 處理，敘述性統計、→
3. 特徵工程 (類別變數 dummy, 連續變數標準化 ANOVA 單因子分析、卡方檢定 (Cramer's V)及視覺化(Python)) →
4. 建置迴歸正則化模型(Lasso, Ridge)，使用 Lasso 篩選出的特徵數(剔除不重要的變數)，再使用這些特徵建置 Ridge。



【圖 1】建模流程圖

【表 1】最終建模欄位變數

欄位名稱	中文說明	欄位屬性	資料範例
刷卡筆數(y)	Dependent variable	數字(int64)	1~9999
未出帳單金額(X_1)	尚未出帳金額	數字(int64)	888
刷卡金額(X_2)	刷卡金額(排除負值)	數字(int64)	0~9999999
卡片額度(X_3)	卡片額度	數字(int64)	100000

前一營業日本期帳單已繳金額(X_4)	前一期繳款金額	數字(int64)	777
年齡群組(X_5)	區分成 9 組	文字	未滿 18 歲、36-45 歲
客戶歸屬分行名稱(X_6)	可能與認列業績分行不同	文字	營業部
消費產品地點名稱(X_7)	刷卡地(實際地)	文字	越南
幣別(X_8)	原幣	文字	USD

註.X1~ X4 為 continuous、X5~ X8 為 category

肆、 資料預處理

1. 因資料已於 7 月處理，如：刪除刷卡金額為負數、手續費等一些回扣資料、國外業務手續費資料、多重共線性檢查...等。
2. 另，此次需建模需檢定相關變數共線性不可過高 (>10) 下，清洗某些欄位中資料量較大的歸為一類、較少的資料筆數則與它項合，或依據資料業務實際狀況填補 NaN，如： 消費次數幣別<68 次(0.75 四分位數)的合併

```
(1)final_datav3['客戶歸屬分行名稱'] = final_datav3['客戶歸屬分行名稱'].fillna('未知分行')
(2)final_datav3['幣別'] = final_datav3['幣別'].fillna('TWD')
(3)rare_currency = (final_datav3.groupby(['幣別'])['幣別'].size().sort_values(ascending = False)<
    final_datav3.groupby(['幣別'])['幣別'].size().quantile(0.75)).tail(21)
    final_datav3['幣別'] = final_datav3['幣別'].replace(rare_currency.index, 'Other')
```

【表 2】各幣別刷卡次數

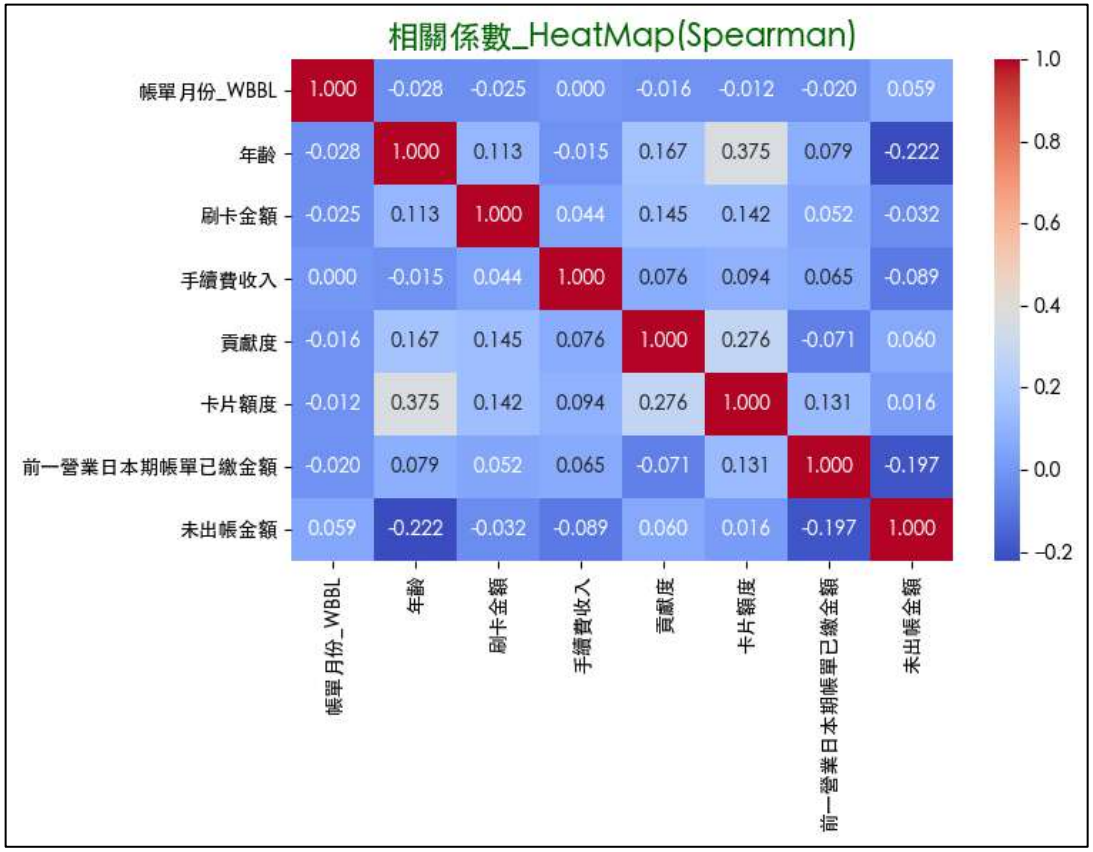
幣別	幣別
TWD	114336
USD	1335
JPY	488
CNY	436
Other	412
AUD	210
EUR	150
KRW	144
HKD	68

伍、連續型相關係數

1. 由於前月未執行常態分布檢定，參數選用 Pearson 相關係數如【圖 1】，導致部分變數間的關係過高，由於刷卡資料高度右邊，白話文講就是多數人刷卡金額及筆數集中在左邊，因此在做連續型變數相關係數時，我們會選擇 Spearman 係數。
2. 此次選擇參數為 Spearman 後，相關係數表如下，最高的為卡片額度與年齡關係為 0.375、依序為卡片額度與貢獻度為 0.276 皆呈低度相關(<0.39)。
3. 另外從表 4 觀察，因各連續變數與刷卡筆數 Spearman 係數相對較高(雖為低度相關)且 p_value≦0.05，因此我們篩選前期帳單已繳金額、未出帳金額、刷卡金額及卡片額度作為連續型顯著變數；後續就目的性選擇模型。
   (1)資料變數多且共線性高使用正則化(Lasso, Ridge)篩選合適變數；
   (2)重解釋刷卡次數因子及異質變異調整，則使用傳統的最小平方法(OLS)；
   (3)需提升預測準確率，則使用機器學習 XGBoost 演算法。

【表 3】Spearman 相關係數

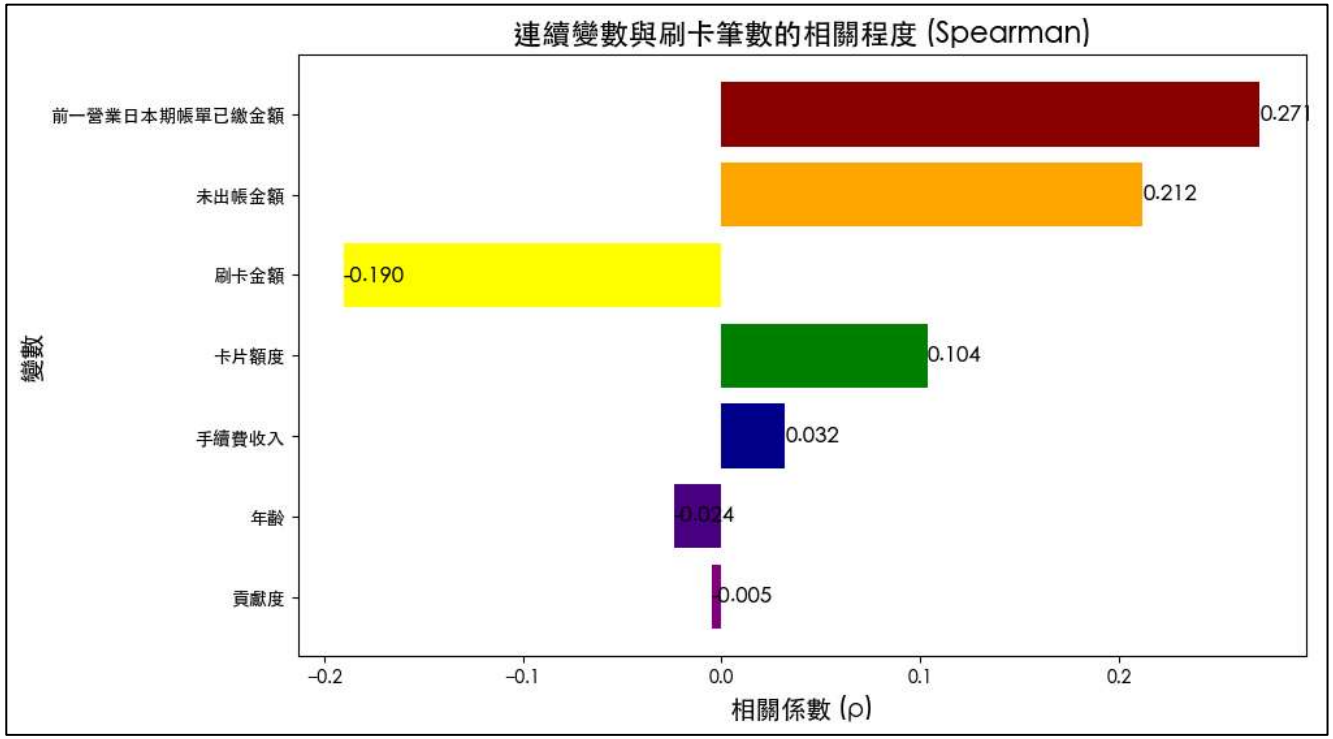
	年齡	刷卡金額	手續費收入	貢獻度	卡片額度	前一營業日本期帳單已繳金額	未出帳金額
年齡	1.0000	0.1134	-0.0151	0.1669	0.3746	0.0789	-0.2222
刷卡金額	0.1134	1.0000	0.0444	0.1451	0.1417	0.0517	-0.0324
手續費收入	-0.0151	0.0444	1.0000	0.0763	0.0940	0.0652	-0.0886
貢獻度	0.1669	0.1451	0.0763	1.0000	0.2762	-0.0711	0.0601
卡片額度	0.3746	0.1417	0.0940	0.2762	1.0000	0.1308	0.0159
前一營業日本期帳單已繳金額	0.0789	0.0517	0.0652	-0.0711	0.1308	1.0000	-0.1973
未出帳金額	-0.2222	-0.0324	-0.0886	0.0601	0.0159	-0.1973	1.0000



【圖 2】Spearman 相關係數

【表 3】與刷卡筆數關係

變數	Spearman 相關係數	p 值	絕對值
前一營業日本期帳單已繳金額	0.2711	0	0.2711
未出帳金額	0.2122	0	0.2122
刷卡金額	-0.1901	0	0.1901
卡片額度	0.1036	1E-277	0.1036
手續費收入	0.0321	3.64E-28	0.0321
年齡	-0.0240	1.89E-16	0.0240
貢獻度	-0.0049	0.0925	0.0049



【圖 3】Spearman 相關係數熱圖

陸、單因子 ANOVA 分析找顯著變數

1. 篩選顯著變數邏輯為：(a)選擇解釋力高變數、(b)剔除關係係數過高變數。
2. 故透過【表 1、表 2】選擇別變數解釋力較大且剔除 Cramer's V 係數過高的變數，避免導致共線性過高，從下表篩選出類別型顯著變數依序為：**客戶歸屬分行名稱、消費產品地點名稱、幣別及年齡群組**。

【表 4】類別變數解釋能力

變數	平方和(SS)	自由度(DF)	均方(MS)	F 統計量	P_value(p-unc)	解釋力(np2)	意義
客戶歸屬分行名稱	11436108.04	35	326745.944	1583.68821	0	0.3212236	Large
認列業績分行名稱	10041403.43	34	295335.395	1357.92582	0	0.28201391	
消費產品地點名稱	3077885.885	20	153894.294	556.1796	0	0.08644276	

幣別	242358.0254	28	8655.64377	30.9788705	5.7686e-159	0.06206232	
年齡群組	1498802.425	8	187350.303	645.809085	0	0.04209403	Medium
信用卡類別	537868.8335	6	89644.8056	300.549287	0	0.0151061	

【表 7】類別變數關係

Cramer'sV	客戶歸屬分行名稱	認列業績分行名稱	消費產品地點名稱	幣別	年齡群組	信用卡類別	個人評等	客戶風險等級	信用卡等級	國內外結帳地點
客戶歸屬分行名稱		Large	Medium	Medium	Large	Large	Large	Medium	Medium	Small
認列業績分行名稱	Large		Medium	Medium	Large	Large	Medium	Medium	Medium	Small
消費產品地點名稱	Medium	Medium		Large	Medium	Small	Small	Small	Small	Negligible
幣別	Medium	Medium	Large		Large	Medium	Small	Small	Small	Large
年齡群組	Large	Large	Medium	Large		Medium	Medium	Medium	Medium	Small
信用卡類別	Large	Large	Small	Medium	Medium		Medium	Large	Large	Negligible
個人評等	Large	Medium	Small	Small	Medium	Medium		Medium	Small	Negligible
客戶風險等級	Medium	Medium	Small	Small	Medium	Large	Medium		Medium	Negligible
信用卡等級	Medium	Medium	Small	Small	Medium	Large	Small	Medium		Negligible
國內外結帳地點	Small	Small	Negligible	Large	Small	Negligible	Negligible	Negligible	Negligible	
開卡狀態	Small	Small	Negligible	Negligible	Negligible	Negligible	Negligible	Negligible	Negligible	Negligible
Gender	Medium	Medium	Small	Small	Negligible	Negligible	Small	Small	Negligible	Negligible
星期幾	Negligible	Negligible	Negligible	Small	Negligible	Negligible	Negligible	Negligible	Negligible	Negligible

註. <0.01Negligible、<0.06Small、<0.14Medium、else Large

柒、建立正則化模型(Lasso, Ridge)

1. 定義：讓機器學習模型在保留多個變數、權重平均分攤減少訓練誤差的前提下，降低共線性避免產生過度配適。使用一個參數λ在訓練過程中對係數向量壓抑以防止係數過大。
2. Ridge 最小損失函數(MSE)並已向量型式表示：

$$\hat{y}^{train} = h(x; w) = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n \quad (1)$$

$$J(w) = \frac{1}{m}(\| Xw - y \|^2 + \lambda w^{Tw}) \quad (2)$$

其中$J(w)$為向量w的損失函數

h方法是計算 $y^{train} - \hat{y}^{train}$ 的均方誤差(MSE)

λ為控制變數

w為係數向量

捌、結論與建議

【表 5】不同套件正則化比較表

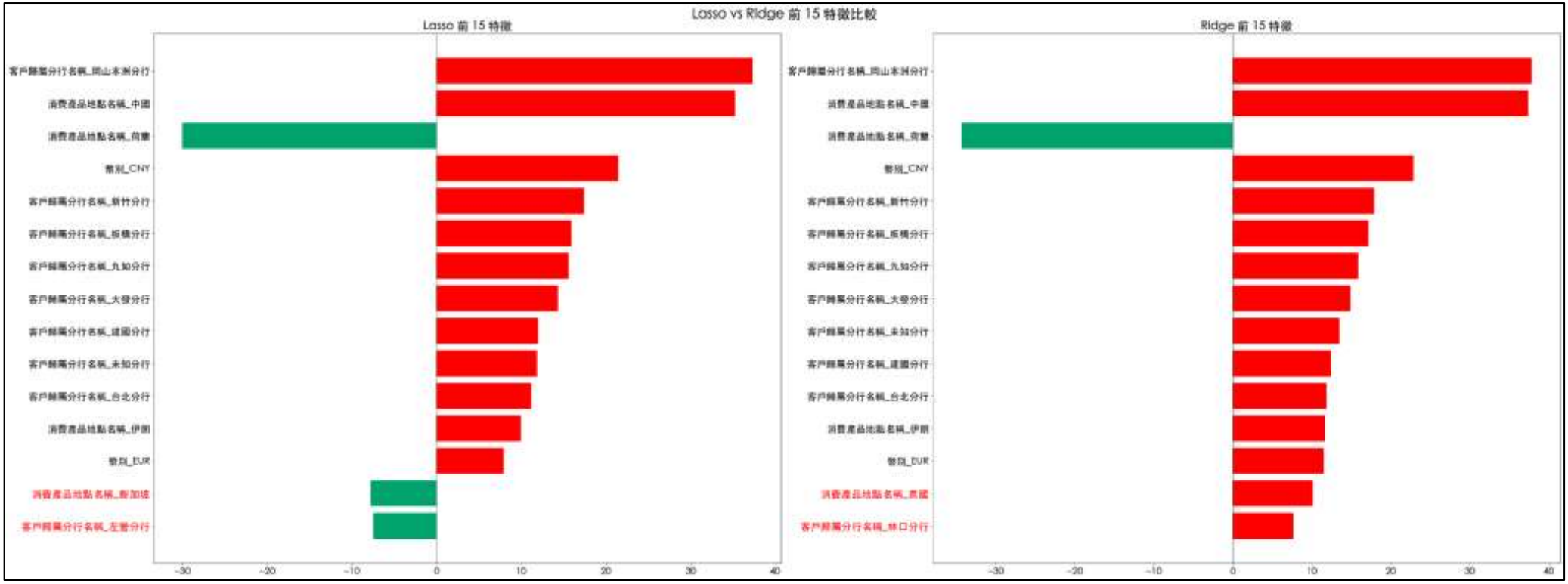
	alpha_sklearn	alpha_正規化	訓練集 Adj R^2 _sklearn	訓練集 Adj R^2 _正規化	測試集 Adj R^2_sklearn	測試集 Adj R^2_正規化
Lasso	0.0047	0.0047	0.4415	0.4359	0.4213	0.4222
Ridge	0.1748	0.1322	0.4421	0.4424	0.4219	0.4287

註.α通常落在 0.1~10 之間 (經驗法則，大部分上夜/社會科學資料集範圍)

1. 此次商業問題是想探討影響消費者刷卡動機，初步共計 49 個變數納入模型，透過資料清洗及正則化選模後，僅剩 8 個顯著變數最後透過綜合變數標準化及 dummy(拆分後變數皆保留以便解釋)後，使用正則化過程(Lasso 篩選後, Ridge 套用特徵)及 sklearn(OneHotEncoder)方式來篩選最適模型。
2. 從【表 8】可以觀察到使用兩種方式建置正則化模型，不論 Lasso (做降維 + 特徵選擇)或 Ridge(特徵上建模)，模型解釋能力差差異不大；另外從兩種模型來比較，可以看出兩模型不論在 alpha, Adj R^2都差異不大。
3. 此次商業問題著重解釋影響刷卡因子的變數而非模型準確率，故變數是否顯著能夠解釋來得比準確更為重要，如：年齡偏低者刷卡意願較年齡高者、行員容易週末刷卡，部分地區刷卡金額不高但刷卡筆數多高。未來除了可以針對不同分行客群特性，使用不同行銷方式，如：百貨駐

點、異業結盟回饋。

4. 另外可針對不同商業問題，使用不同的統計模型及方式來擬定相關策略，如：沒有信用卡客戶是否會辦法使用邏輯斯迴歸；如果是想要預測客戶刷卡準確率可使用機器學習，如：k_means, XGBoost。
5. 最終 Ridge 模型篩選前 15 個 Ridge Equation 及重要變數如下：可以觀察到 Ridge 相對於 Lasso 模型可以將細數 shrink 至最小，但仍會保留變數；而 Lasso 則是將係數壓縮至 0 後，接續篩選較為重要之變數



【圖 4】Lasso& Ridge 前 15 項特徵比較

1. 6. Lasso 迴歸方程式

Ŷ = 7.0731 +
37.3390 * 客戶歸屬分行名稱_岡山本洲分行 +
35.2601 * 消費產品地點名稱_中國 +
-29.9845 * 消費產品地點名稱_荷蘭 +
21.4637 * 幣別_CNY +
17.4380 * 客戶歸屬分行名稱_新竹分行 +
15.9006 * 客戶歸屬分行名稱_板橋分行 +
15.5759 * 客戶歸屬分行名稱_九如分行 +
14.3429 * 客戶歸屬分行名稱_大發分行 +
12.0112 * 客戶歸屬分行名稱_建國分行 +
11.8473 * 客戶歸屬分行名稱_未知分行 +
11.2213 * 客戶歸屬分行名稱_台北分行 +
9.9418 * 消費產品地點名稱_伊朗 +
7.9548 * 幣別_EUR +
-7.7557 * 消費產品地點名稱_新加坡 +
-7.4977 * 客戶歸屬分行名稱_左營分行

7. Ridge 迴歸方程式

Ŷ = 5.2569 +
37.8387 * 客戶歸屬分行名稱_岡山本洲分行 +
37.3943 * 消費產品地點名稱_中國 +
-34.3741 * 消費產品地點名稱_荷蘭 +
22.8400 * 幣別_CNY +
17.8712 * 客戶歸屬分行名稱_新竹分行 +
17.1363 * 客戶歸屬分行名稱_板橋分行 +
15.8386 * 客戶歸屬分行名稱_九如分行 +
14.8870 * 客戶歸屬分行名稱_大發分行 +
13.4786 * 客戶歸屬分行名稱_未知分行 +
12.3785 * 客戶歸屬分行名稱_建國分行 +
11.8396 * 客戶歸屬分行名稱_台北分行 +
11.6761 * 消費產品地點名稱_伊朗 +
11.4959 * 幣別_EUR +
10.1190 * 消費產品地點名稱_英國 +
7.6272 * 客戶歸屬分行名稱_林口分行

玖、附件

殘差檢定是否符合常態分布、同質變異、獨立性，由於信用卡資料涉及到時間序列，交易金額大的交易筆數與交易金額小的交易筆數會變異較大，因此造成殘差存在異質變異，若今天是為了統計上的模型嚴謹程度，那就得繼續去修改係數或轉換變數，如：最小加權平方(WLS)、Box-Cox..等