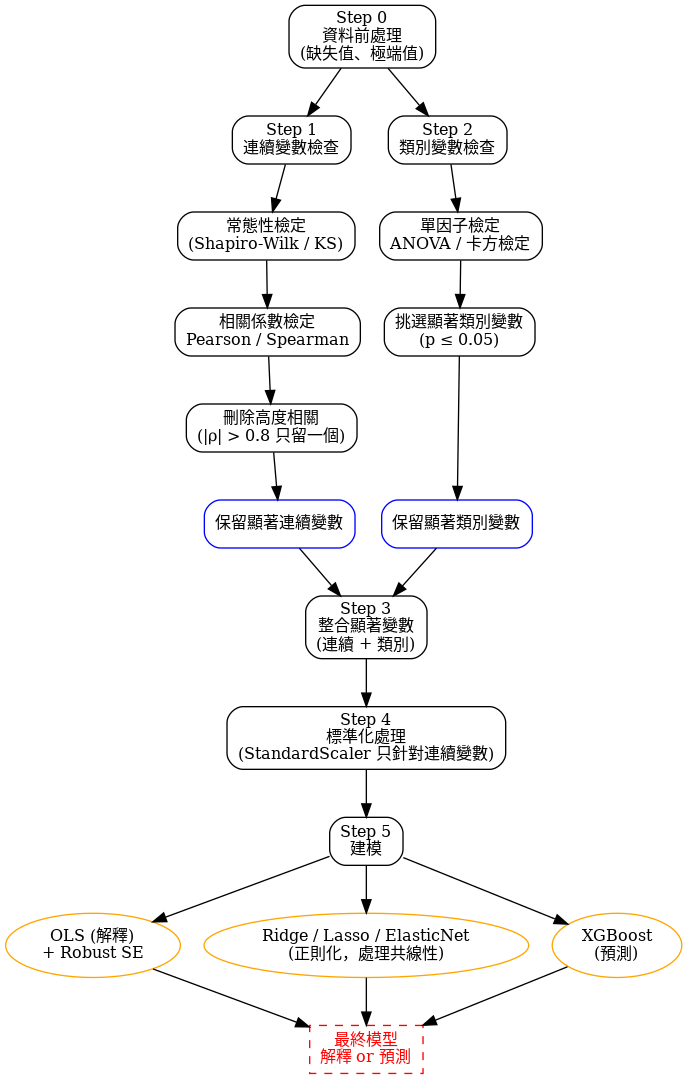
**信用卡客戶統計建模**

1. **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。
2. **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)。

1. **建模過程及**

查找資料源(Tableau) 🡪 資料前處理，如單因子分析(ANOVA)及視覺化(Python) 🡪 特徵工程 (類別：dummy,連續：標準化) 🡪 建立迴歸模型(傳統OLS, Logistic Regression, 機器學習XGBoost)



【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(X1) | 尚未出帳金額 | 數字(int64) | 888 |
| 刷卡金額(X2) | 刷卡金額(排除負值) | 數字(int64) | 0~9999999 |
| 卡片額度(X3) | 卡片額度 | 數字(int64) | 100000 |
| 前一營業日本期帳單已繳金額(X4) | 前一期繳款金額 | 數字(int64) | 777 |
| 年齡群組(X5) | 區分成9組 | 文字 | 未滿18歲、36-45歲 |
| 客戶歸屬分行名稱(X6) | 可能與認列業績分行不同 | 文字 | 營業部 |
| 消費產品地點名稱(X7) | 刷卡地(實際地) | 文字 | 越南 |
| 幣別(X8) | 原幣 | 文字 | USD |

註.X1~ X4為continuous、X5~ X8為category

1. **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…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')

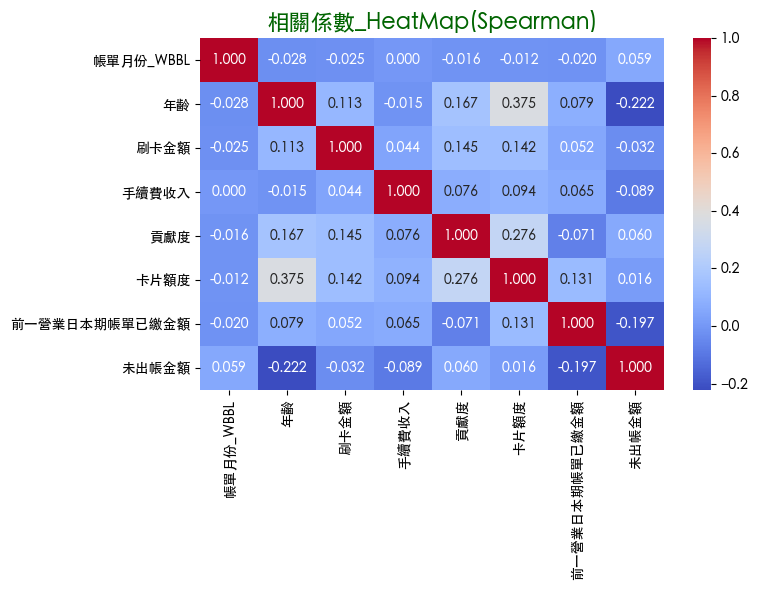
【表2】各幣別刷卡次數

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

1. **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4】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5】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 | 0.271057 | 0 | 0.271057 |
| 未出帳金額 | 0.212179 | 0 | 0.212179 |
| 刷卡金額 | -0.19011 | 0 | 0.190112 |
| 卡片額度 | 0.103568 | 1E-277 | 0.103568 |
| 手續費收入 | 0.032085 | 3.64E-28 | 0.032085 |
| 年齡 | -0.02399 | 1.89E-16 | 0.023994 |
| 貢獻度 | -0.00491 | 0.09247 | 0.004907 |

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

AI 產生的內容可能不正確。

【圖1】Spearman相關係數熱圖

1. **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6】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

1. **建立正則化模型(Lasso, Ridge)**
2. **結論與建議**

【表8】正則化比較表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | alpha\_sklearn | alpha\_正規化 | 訓練集 Adj R² \_sklearn | 訓練集 Adj R² \_正規化 | 測試集 Adj R²\_sklearn | 測試集 Adj R²\_正規化 |
| Lasso | 0.00465 | 0.00469 | 0.441361 | 0.435930 | 0.4210844 | 0.422162 |
| Ridge | 0.17475 | 0.13219 | 0.441899 | 0.442411 | 0.4217084 | 0.428684 |

* 1. 因我這邊商業問題是想探討影響消費者刷卡動機，故選擇迴歸模型中的Lasso, Ridge兩模型來消除共線性過高的狀況
  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Ridge模型不論在alpha, Adj R²都略優於Lasso模型，
  3. 此次定義問題是重解釋而非準確率，故變數是否顯著來得比準確更為重要。未來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。