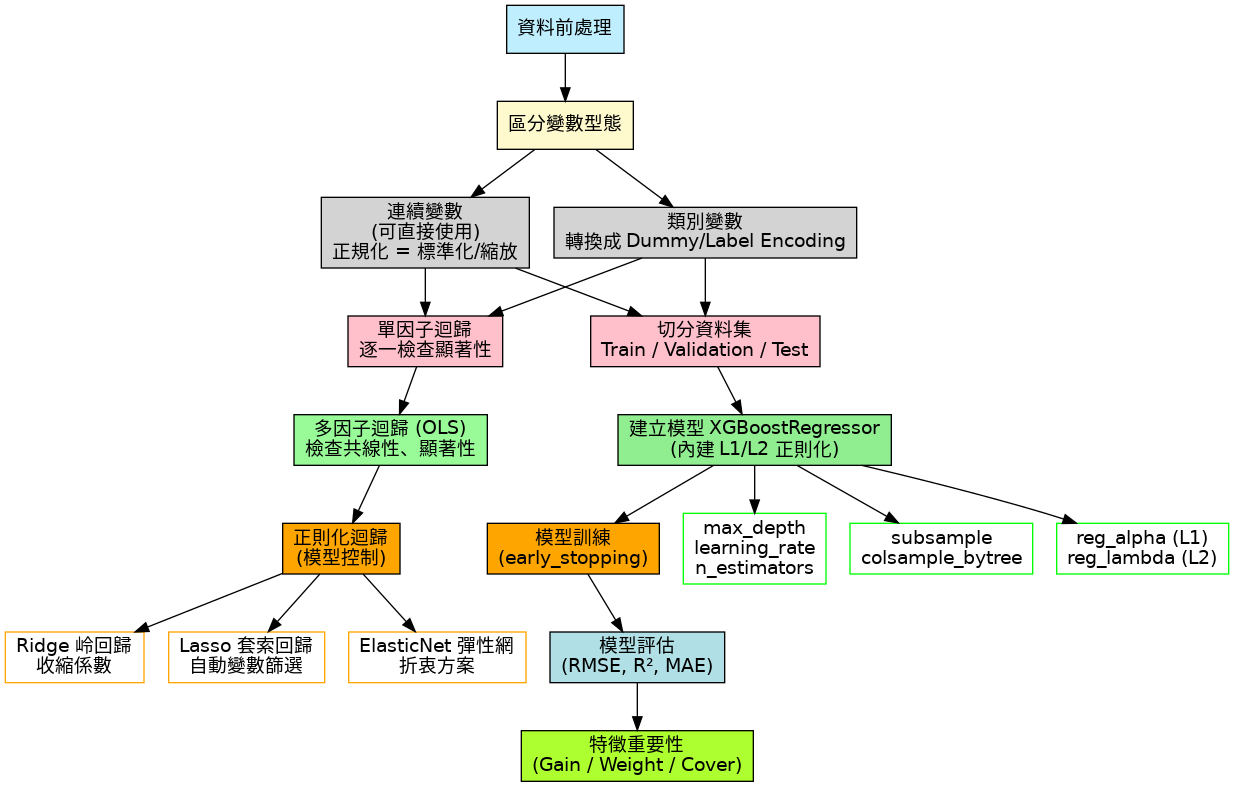
**信用卡客戶統計建模**

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. 可以針對刷卡筆數高之分行且那幾天，提供刷卡回饋較多活動來吸引客群，透過客戶介紹客戶也能更快速累積
   4. 後續欲持續針對其它類別變數作為欄位資訊納入分析資料，並將相關報表的日期欄位釐清更清楚，以便進一步建立模型探討影響本行客戶刷卡意願相關因子。
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**

查找資料源(Tableau) 🡪 資料前處理及視覺化(Python) 🡪 特徵工程(單因子ANOVA、dummy, 正則化) (Lasso, Ridge) 🡪 建立迴歸模型(傳統OLS, Logistic Regression, 機器學習XGBoost)



【表1】資料欄位

|  |  |  |  |
| --- | --- | --- | --- |
| **欄位名稱** | **中文說明** | **欄位屬性** | **資料範例** |
| 刷卡金額 | 刷卡金額(排除負值) | 數字(int64) | 0~9999999 |
| 未出帳單金額 | 尚未出帳金額 | 數字(int64) | 888 |
| 前一營業日本期帳單已繳金額 | 前一期繳款金額 | 數字(int64) | 777 |
| 年齡群組(Category) | 區分成9組 | 文字 | 未滿18歲、36-45歲 |
| 客戶歸屬分行名稱(Category) | 可能與認列業績分行不同 | 文字 | 營業部 |
| 消費產品地點名稱(Category) | 刷卡地(實際地) | 文字 | 越南 |
| 卡片張數 | 較客戶數多(排除重複) | 數字(int64) | 1 |
| 幣別(Category) | 原幣 | 文字 | USD |
| 刷卡筆數 | 刷卡筆數 | 數字(int64) | 1~9999 |
| 刷卡金額 | 刷卡金額(排除負值) | 數字(int64) | 0~9999999 |
| 手續費收入 | 手續費(有負值) | 數字(int64) | -114~95542 |
| 貢獻度 | 信用卡貢獻業績(有負值) | 數字(int64) | -7665~283468 |
| 卡片額度 | 卡片額度 | 數字(int64) | 100000 |

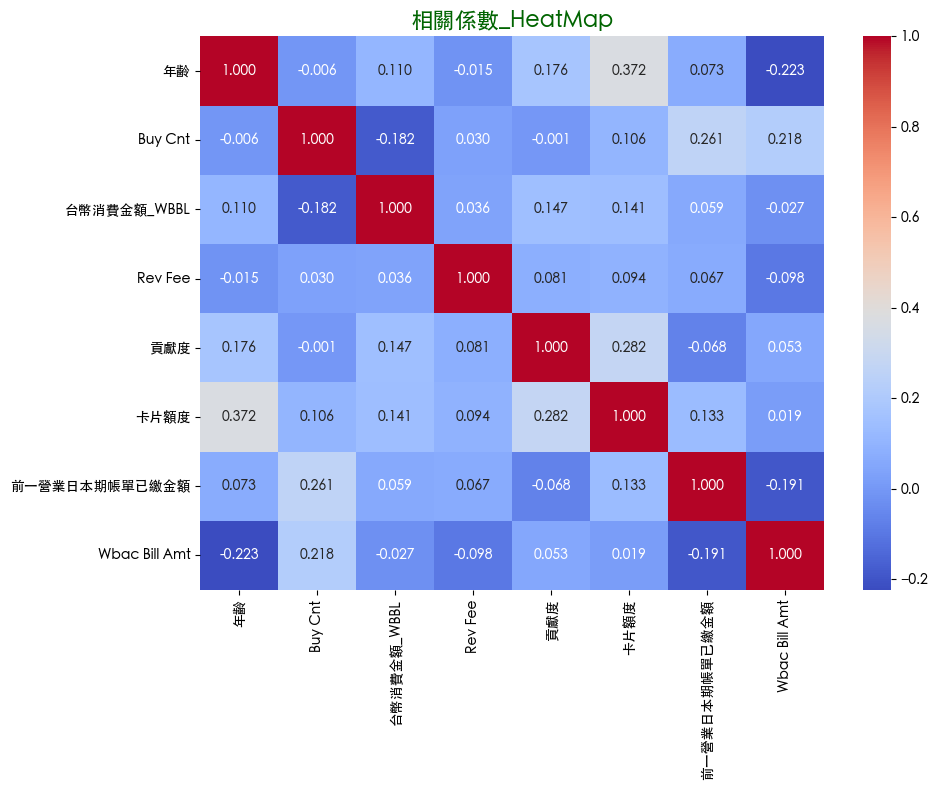
【表2】敘述性統計

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| index | 年齡 | 刷卡筆數 | 刷卡金額 | 手續費 | 貢獻度 | 卡片額度 | 本期帳單已繳金額 | 尚未出帳金額 |
| mean | 52 | 12 | 1,811 | 269 | 4,260 | 152,038 | 12,175 | 10,698 |
| std | 13 | 17 | 7,552 | 4,226 | 14,102 | 157,534 | 33,338 | 27,357 |
| min | 19 | 1 | 1 | -114 | -7,665 | 20,000 | 0 | 0 |
| 25% | 42 | 2 | 192 | 0 | 9 | 100,000 | 0 | 502 |
| 50% | 55 | 5 | 500 | 0 | 390 | 100,000 | 184 | 3,371 |
| 75% | 62 | 13 | 1,302 | 0 | 3,111 | 200,000 | 11,012 | 11,186 |
| max | 102 | 98 | 1,000,000 | 95,542 | 283,468 | 2,000,000 | 370,000 | 438,254 |

1. **資料預處理**
   1. 因資料已於7月處理，如：刪除金額為負數、手續費等一些回扣資料、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併等等。
2. **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相關係數熱圖

【表4】連續變數與刷卡筆數關係

|  |  |  |  |
| --- | --- | --- | --- |
| 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1】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2】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. **雙因子ANOVA分析找顯著變數**
2. **結論與建議**
   1. 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、屏東兩南部分行外，其餘皆為台中以北之分行，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。
   2. 可以針對刷卡筆數高之分行且那幾天，提供刷卡回饋較多活動來吸引客群，透過客戶介紹客戶也能更快速累積
   3. 後續欲持續針對其它類別變數作為欄位資訊納入分析資料，並將相關報表的日期欄位釐清更清楚，以便進一步透過分群模型預測本行無申辦信用卡客戶的辦卡的機率。
3. 變數相關性
   1. 客戶歸屬分行名稱與認列業績分行名稱相關性高，因此選擇解釋性較高的歸屬分行
   2. 客戶風險等級、信用卡等級與信用卡類別相關性高，因此我選擇信用卡類別作爲顯著變數；
   3. 另，檢視部分消費地點資料可能是在網路上，僅因爲產品是在國外，與實際場域有所出入，因此選擇幣別作為顯著變數；
   4. 國內外結帳地點為我自行下條件(國外金額為0或國內外金額相等)，視為國內其餘國外亦納入顯著變數
   5. 剩餘變數為