信用卡違約預測

組員：林貫原、許政揚、周昱宏、楊廷紳、易祐辰、留筠雅

June 17, 2024

隨著現代社會的發展，信用卡已成為人們日常生活中不可或缺的支付工具之一，但與此同時，信用卡違約的問題也逐漸浮現。信用卡違約指的是持卡人未能按時或完全償還信用卡欠款的情況，這可能導致嚴重的財務後果，不僅對持卡人自身造成負擔，還可能對金融系統穩定產生影響。研究信用卡違約的統計，有助於深入了解違約行為的特徵、趨勢和影響因素，進而提供預測和管理違約風險的依據。這方面的研究對於金融機構評估信用風險、制定信用政策以及發展適當的風險管理策略至關重要。此外，對於消費者而言，了解違約的可能原因和風險因素，可以引導其更加理性地使用信用卡，避免陷入財務困境。

信用卡違約研究的背景包括但不限於以下幾個方面：首先，需考慮宏觀經濟環境的影響，如經濟增長率、失業率、通脹率等因素對信用卡違約率的影響。其次，個人特徵和行為模式也是影響違約的重要因素，如持卡人的年齡、收入水平、職業、婚姻狀況等。此外，信用卡產品本身的特性、使用方式以及支付習慣也會對違約率產生影響。最後，法律法規和監管政策對信用卡市場的規範也將對違約情況產生一定的影響。

1 研究目的

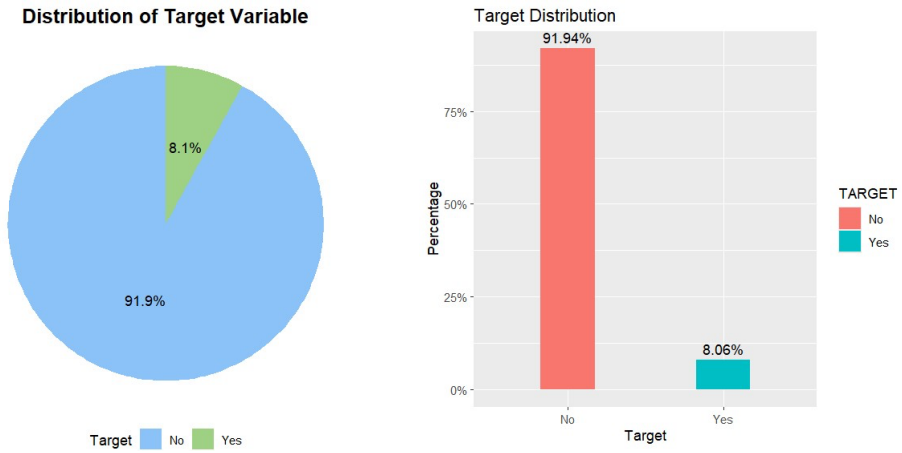
本研究旨在以客戶基本資料與交易資料為基礎，探討信用卡違約風險的判斷方法。具體而言，我們將透過以下步驟來實現研究目標：首先，挖掘資料中有用的變數，利用特徵選取等方法，對資料進行分群，以識別出對信用卡違約具有重要影響的變數。其次，找出違約客戶的分群特徵，透過分群分析，探索和理解違約客戶的特徵和行為模式，以協助金融機構更好地理解和管理風險。最後，建立分群前與分群後的預測模型，並進行比較分析，以評估分群對信用卡違約預測的影響，從而提高預測準確性。此外，我們將開發 R Shiny 網頁 demo，用

於展示已完成的分析結果，以提供更直觀和易於理解的方式來呈現研究成果。透過以上目標的達成，我們期望能夠為金融機構提供更準確的信用卡違約風險評估模型，促進風險管理的效率和效果，同時為客戶提供更安全可靠的金融服務。

2 資料介紹

此資料集是班加羅爾國際資訊科技學院 (International Institute of Information Technology Bangalore) 收集而來的。共有 122 個變數。分為訓練集與測試集，訓練集資料共有 306611 筆，並無重複紀錄的情形，測試集資料是從訓練集資料中抽取，共有 900 筆。

此資料為不平衡資料，違約的資料筆數為 24835，正常使用信用卡的資料筆數為 281776，其比例如圖 1 ，此種資料若未做處理直接分析的話，我們的模型會大量的學習到正樣本的資料，在預測時很容易發生過度配適 (Overfitting) 的問題，假設正常比違約的比例是 99:1，我們的模型預測 100 份資料都是正常，這樣的準確率是 99%，可是他完全沒有預測到錯誤的那筆，那這個模型似乎用處不是特別大。因此在做分析之前須先對不平衡資料做處理，詳細處理方式將於下一小節做說明。



(a) 圓餅圖　　(b) 長條圖

圖 1: 不平衡資料示意圖

因為變數過多，本小節並不會一一敘述每個變數，若想詳細了解每個變數可以參考附錄。

3 資料處理

在做資料分析之前需要先進行資料處理，這是因為資料集中可能存在各種問題，如遺失值、異常值等，這些問題會影響分析的準確性和可信度。本小節會詳細說明我們對此資料做了什麼處理，並解釋原因。

3.1 去除遺失值過多變數

觀察資料後，發現資料集中有許多變數存在大量遺失值。這種情況可能使分析結果出現偏差，因此決定對遺失值超過 32% 的變數進行刪除。原因如下：

- 如果變數中大部分的值都是遺失值，即使使用插補方法填補缺失值，也難以保證填補後的資料能夠完全準確地反映真實情況。
- 圖 2 將所有具有遺失值的變數按其遺失率排序，顯示部分變數的遺失值比例超過了 49%。這種高比例的缺失值可能對模型的建構和預測產生不利影響，即使使用插補方法也難以取得良好效果。儘管變數 `OCCUPATION_TYPE` 的遺失率為 31% 並不低，但考慮到該變數可能對後續分析具有重要性，我們選擇保留該變數，並以 32% 作為刪除的界線。
- 我們認為如果該變數是重要變數，那公司應該會特別要求客戶盡可能填寫該資料，因此若遺失值太多，或許也代表該公司並不是很在乎該變數。

綜上所述，我們的刪除決策旨在確保模型建構和預測的準確性，同時保留重要變數以支持後續的分析工作。被刪除變數詳細的資訊可參考附錄表 9，附錄圖 31 為刪除變數按遺失率高低排序的長條圖。

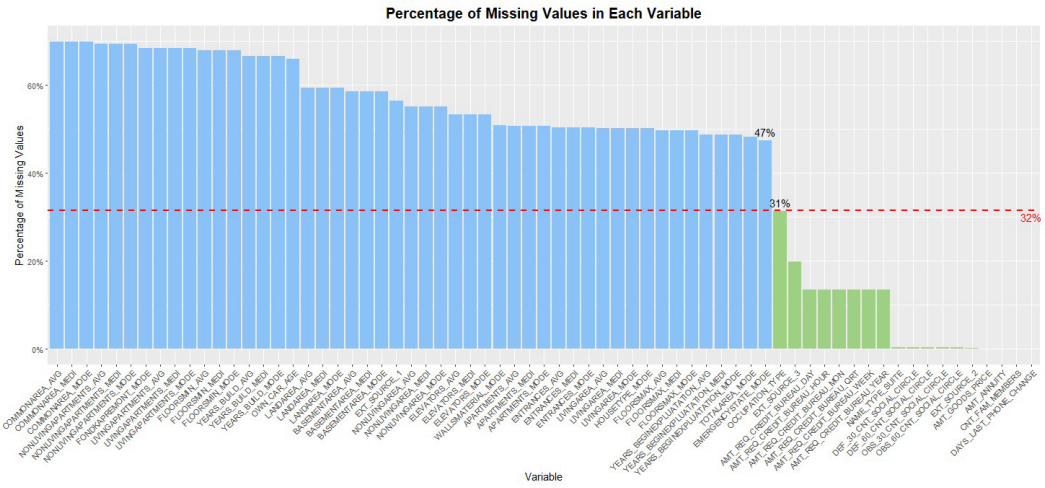


圖 2: 有遺失值變數的遺失率

3.2 調整變數

觀察資料後，我們注意到一些變數或許可以進行調整以更好地進行分析探討。因此，我們新增了以下新變數：

1. SUM_FLAG_DOCUMENT：客戶總共簽署的文件數量。

   我們發現此資料集包含有關客戶是否簽署文件的相關資訊。這些文件共有二十種不同的類型，然而其具體內容因可能涉及商業機密而無法提供。儘管如此，我們考慮到這些文件可能具有重要性，故不會將相關變數從資料集中刪除。雖然大多數客戶僅簽署其中一種文件，但也有一部分客戶簽署了多份文件。基於此，我們決定將這二十個文件相關的變數合併為一個新變數，以探討簽署文件的數量是否與違約情況或其他變數之間存在關係。

2. SUM_AMT_REQ_CREDIT_BUREAU：在申請信用卡之前一年內向信用局查詢客戶信用報告的總次數。

   我們將注意力轉向另外六個變數 (AMT_REQ_CREDIT_BUREAU)。這些變數分別探討了在申請信用卡之前一個小時內向信用局查詢客戶信用報告的次數、一天內 (不包括前一個小時)、一星期 (不包括前一天)、一個月 (不包括前一星期)、一季 (不包括前一個月)、以及一年 (不包括前一季) 的次數。在進行分析之前，我們推測這些變數中可能只有其中幾個對我們的目標變數具有一定相關性。期望透過結合這些變數，提高對目標變數相關性的解釋力。我們已將這些變數的意義以示意圖的形式呈現，詳見圖 3 ，有助於我們更好地理解變數間的關係，以及整體的趨勢。

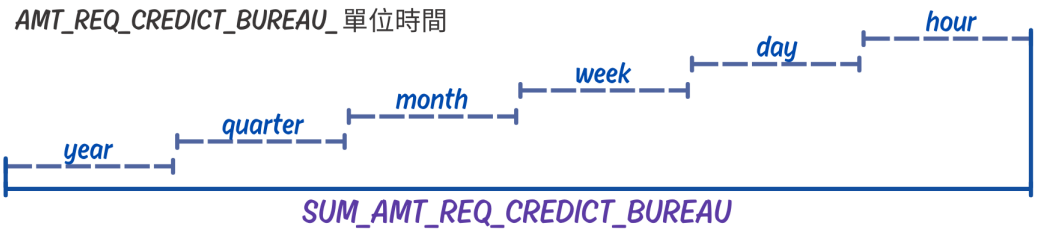
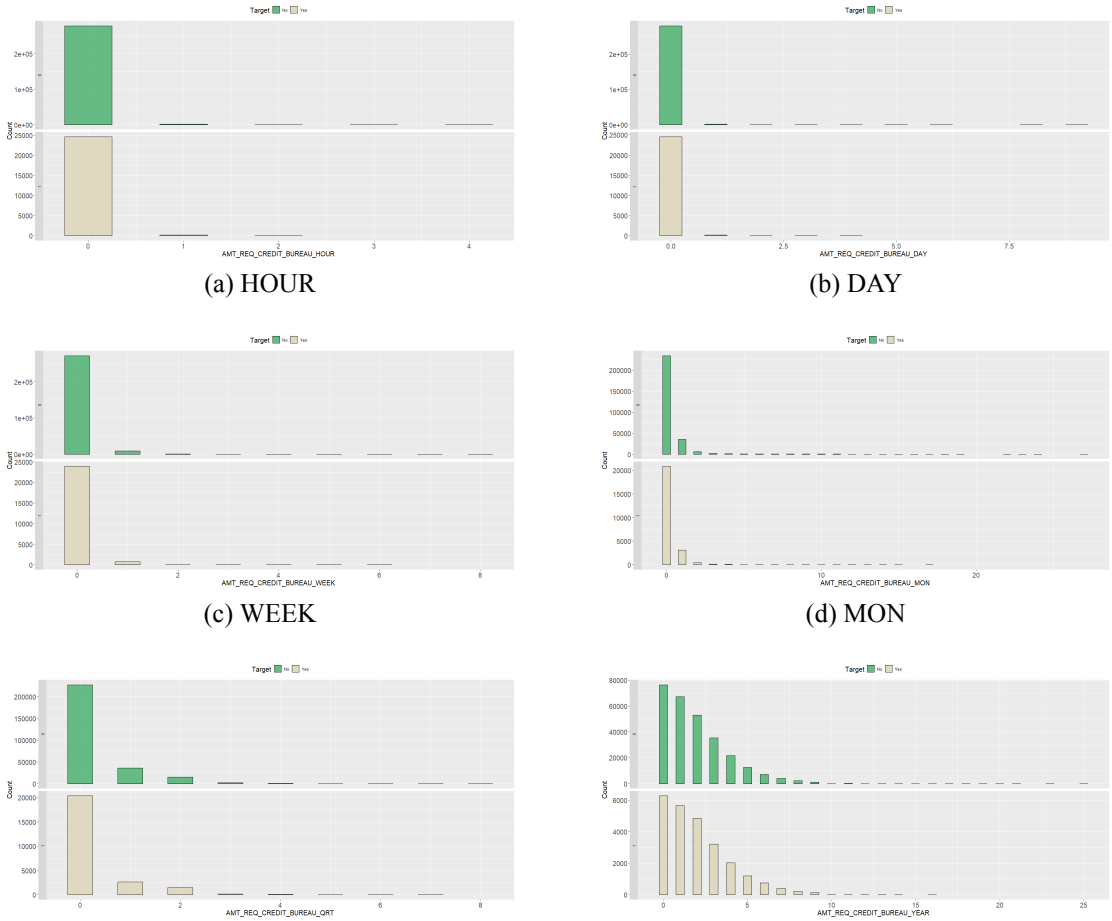


圖 3: 變數 AMT 示意圖

圖 4 展示了 AMT 系列的六個變數長條圖，並根據我們的目標變數進行分組，綠色表示無違約情形 (TARGET=0)，淺土黃色為有違約情形 (TARGET=1)，從圖中可以觀察到，這些變數的分佈不會因有無違約而有

所改變。因此，在做相加後的新變數分佈也會與原始數據的分佈相同，不會有任何問題。



(a) HOUR (b) DAY

(c) WEEK (d) MON

(e) QRT (f) YEAR

圖 4: 變數 AMT 長條圖

3. `missing_ratio`：該筆資料中的遺失值個數佔刪除了超過 32% 遺失值變數後全部變數的比例。

   在進行分析之前，我們考慮到可能存在一些信用卡違約的個案可能不太願意填寫所有的資料，因此我們希望通過這個變數來探討遺失值的比例與違約情況之間的關係。

變數資料修改：

1. 有五個變數依序在探討客戶申請信用卡時的年齡、申請信用卡前多少天開始目前的工作、申請信用卡前多少天更改了註冊資料、申請貸款前多少天更改了申請貸款的身份證件、申請信用前多少天換手機，單位皆為天數，且是由調查當天往前計算，因此值皆為負數，為了後續分析方便，我們將

所有資料除以 365.25，讓單位由天改為年，並取絕對值。最後再將變數名稱從原先的 DAY 開頭改為 YEAR 開頭，當然變數的含義也會有所不同，比如說申請信用卡前多少天開始目前的工作就變為申請信用卡前從事目前工作的年資。

2. 變數 GENDER 中的 XNA 改為新類別 others。考慮到當前社會中性別平等意識的興起，雙性戀或同性戀的人數逐漸增多，因此可能有一些客戶認同的性別身份不符合傳統二元性別觀念。這些遺失值可能代表著個案不願意填寫性別資訊，因為他們可能是男性但認同為女性，或是女性但認同為男性。考慮到這種情況的存在，我們決定不對遺失值進行插補，而是將其歸類為一個新的類別，以更好地反映多元性別認同。

3. 變數 ORGANIZATION_TYPE 中的 XNA 改為新類別 Pensioner，表示其為退休的人，原因為對應到 YEAR_EMPLOYED ，其值皆為 1000，該變數是探討在申請信用卡前從事目前的工作多少年，考慮兩者同時出現此異常的情形，我們推測是因為受訪者已退休而無工作，才會有這樣的紀錄。

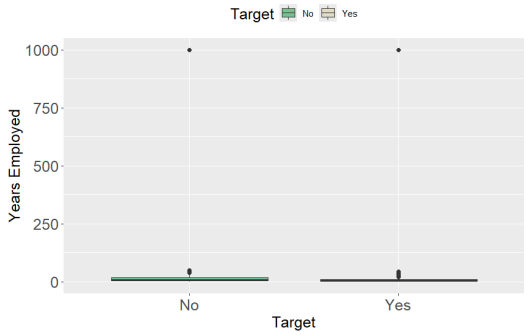
4. 變數 ORGANIZATION_TYPE 用以描述客戶所屬的公司類型。由於該變數涉及的公司類型共有 58 種，數量較多，可能會對後續分析產生一定的影響。因此，在此將一些相似的公司類型進行了合併。具體而言，將 Business Entity Type 1 到 Business Entity Type 3 合併為 Business Entity Type，將 Industry: type 1 到 Industry: type 13 合併為 Industry，將 Trade: type 1 到 Trade: type 7 合併為 Trade，將 Transport: type 1 到 Transport: type 4 合併為 Transport，而其餘類型則保留原始的分類。最終，將原先的 58 種不同公司類型縮減為 35 種。

3.3 離群值處理

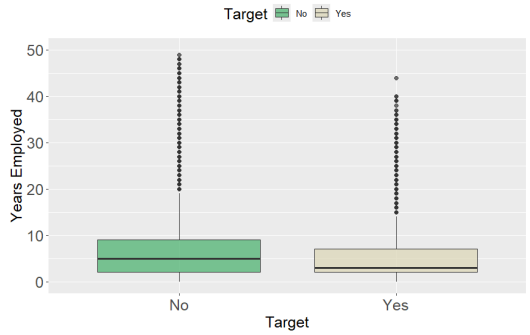
當我們調整了變數後，接著觀察每個變數的資料分佈時，我們發現了許多變數都有出現離群值的情況。然而，在深入探討具有離群值的變數意義後，我們決定不對這些離群值進行特別處理。因為基於對資料的真實性和完整性的考慮且這些資料是真實收集而來的，因此我們認為這些離群值可能代表了真實世界中的特殊情況，而非收集時的錯誤或異常。因此保留離群值較能反應真實的數據，除了幫助我們更全面地理解資料，更協助找出可能存在的特定模式或趨勢。

需要特別注意的是，在申請信用卡前多少天開始目前的工 作的個變數中，有很多資料顯示天數為-365243，經過轉換過後大約是 1000 年，相當不符合常理，但看到職業類別是顯示已退休（原為 XNA，經過變數調整），我們判斷是訪問者故意設定的數值，經過思考後，我們覺得退休後雖然還是有可能會從事工作，但該變數是指此前工作的天數，若有工作應該也會填上天數，因此我們判斷這些資料是已退休且沒工作的，不過若直接將這群資料皆轉換成 0 天的話，那個 0 是會有意義的，像是剛出社會的新鮮人尚未找到工作，那也會是顯示 0 天，因此我們也不想直接將這些資料改為 0，將保留原本的資料待後面分析去做分群討論，查看是否退休的人比尚未退休的人更容易有信用卡違約的情形發生。

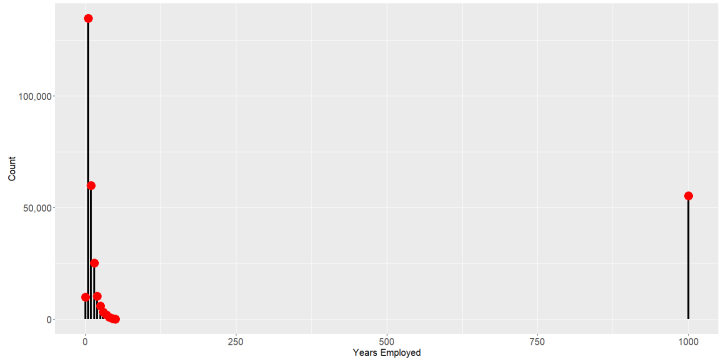
為了更好說明該變數的所有資料中具有異常值，詳見圖 5 ，(a) 為保留 1000 年資料的直方圖，可以觀察到它與其他資料距離非常遠；(b) 為去除 1000 年資料的直方圖，發現若沒有這些資料，雖然依舊有離群值的出現，但也蠻合理的，畢竟一直待在同一間公司工作要超過 20 年的人本來就會很少；(c) 為整體以五年為一個組距後的莖葉圖，明顯可以說明年數中間有個斷層，代表這群資料確實是異常值，該做探討。



(a) 直方圖（保留一千年）



(b) 直方圖（去除一千年）



(c) 莖葉圖

圖 5: 申請信用卡前該工作年資

3.4 遺失值處理

為了更好的面對遺失值的處理問題，以下我們將依處理方式來做探討：

1. 平均數插補（連續型變數）：

   - `AMT_ANNUITY` (12 筆)：

     這個變數代表客戶的貸款年金。其敍述統計量如表 1 ，因其遺失率較低，我們決定將這些遺失值替換為平均數 27123.36。

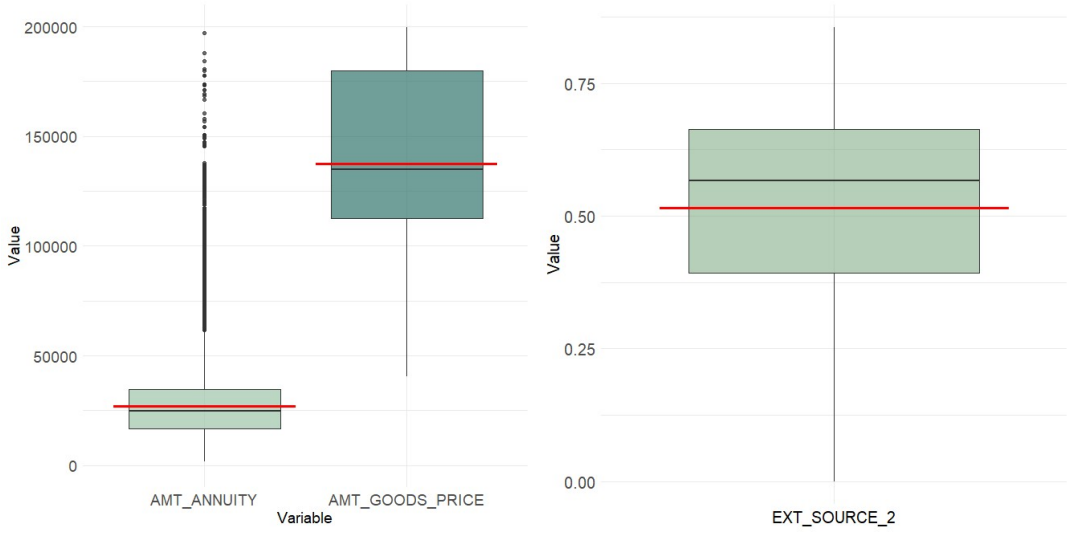
   - `AMT_GOODS_PRICE` (276 筆，佔全部資料的 0.09%)：

     這個變數代表客戶欲購買商品的價格。其敍述統計量如表 1 ，因其遺失率較低，我們決定將這些遺失值替換為平均數 538694.10。

   - `EXT_SOURCE_2` (658 筆，佔全部變數 0.2%)：

     這個變數代表外部數據來源 2 的正規化分數。其敍述統計量如表 1 ，因其遺失率較低，我們決定將這些遺失值替換為平均數 0.5143。

   圖 6 為三個變數的箱型圖，可從圖中觀察到三者平均數和中位數非常接近，通常表示數據呈現對稱分佈或接近對稱分佈，這意味著數據的中心位置相對固定，數據在中心附近的分佈比較均勻，沒有明顯的偏移或異常值。這種情況下，平均值和中位數都可以作為代表數據集中心位置的指標，而且它們的值很接近，表明數據的集中趨勢相對穩定。



(a) Annuity & Good Price　　(b) EXT2

圖 6: 以平均數插補的變數箱型圖

表 1: 以平均數插補的變數敘述統計

統計量 / 變數	平均數	中位數	標準差	q1	q3
AMT_ANNUITY	27123.36	24930	14475.81	16564.50	34596.00
AMT_GOODS_PRICE	538694.10	450000	369455.07	238500.00	679500.00
EXT_SOURCE_2	0.5143	0.5659	0.1911	0.3924	0.6636

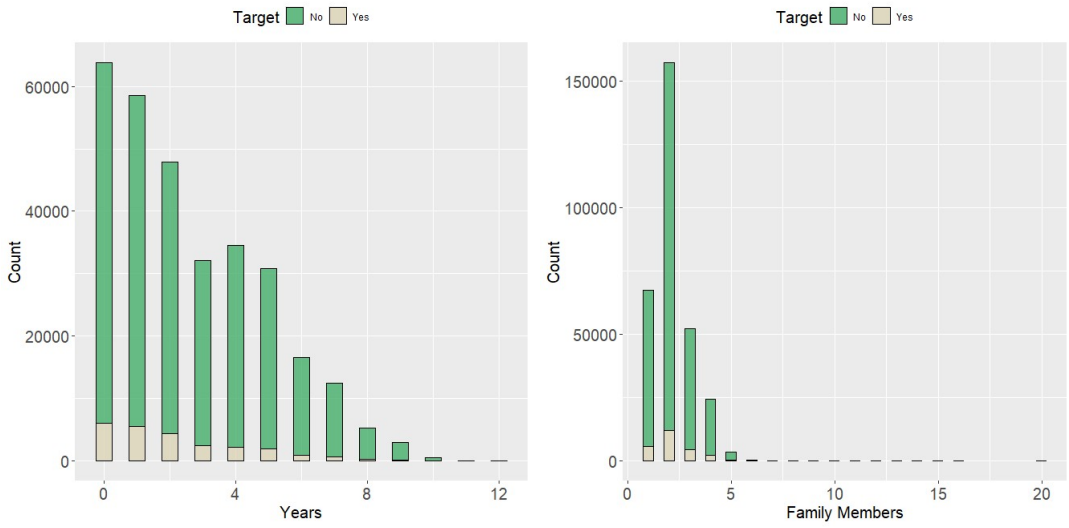
2. 眾數插補：

- YEARS_LAST_PHONE_CHANGE (1 筆)：

  該類別用於記錄客戶在申請貸款前多少年換過手機，在觀察圖 7 後，雖然發現到其分配並沒有明顯差異，但由於只有一筆遺失值，因此我們決定直接使用眾數插補的方法來填補這個遺失值，將這一筆遺失值替換為眾數值 0。

- CNT_FAM_MEMBERS (2 筆)：

  該類別用於記錄客戶家庭成員數量，由於僅有兩筆遺失值，在觀察圖 7 後，發現到客戶中家庭成員數兩員的比例相較其他數量存在明顯差異，因此我們決定直接使用眾數插補的方法來填補這個遺失值，將這兩筆遺失值替換為眾數值 2。



(a) Last Phone Change (b) Members

圖 7: 以眾數插補的變數長條圖

3. 新增新類別：

   - `OCCUPATION_TYPE` (96109 筆，佔全部資料 31%)：

     該類別用於記錄客戶的職業。考慮到大多數人在填寫問卷時會提供自己的職業，因此遺失值很可能表示其他類型的職業，而非填寫時意外遺漏。因此，為了在分析中保留這些資訊，我們決定將所有遺失值歸類為一個新的類別 others。

   - `NAME__TYPE_SUITE` (1289 筆，佔全部資料 0.4%)：

     該類別用於記錄客戶在申請貸款時的陪同人員。由於許多客戶可能沒有在提供的選項中找到適合自己情況的選擇，因此未填寫資料的情況比較常見。我們假設未填寫資料的人可能沒有在選項中找到合適的選擇，而不是因為他們沒有陪同人員。為了將這些未填寫的資料納入考慮，我們將這些遺失值歸類為一個新的類別 Non collected，以表示這部分資料的陪同人員信息是未知的或未提供的。

4. 刪除資料：

   - `OBS_30_CNT_SOCIAL_CIRCLE`、`DEF_30_CNT_SOCIAL_CIRCLE`、`OBS_60_CNT_SOCIAL_CIRCLE`、`DEF_60_CNT_SOCIAL_CIRCLE`

     (1020 筆，佔全部資料的 0.3%)：

     這四個變數屬於相同的類型，並且它們的遺失值同時發生在相同的資料中。儘管這 1020 筆資料中，有 36 筆是屬於違約的部分，佔 1020 筆的 3.5%，與原資料集整體違約率 8.06% 有所差異，但由於遺失值僅佔全部資料的極少部分，因此直接刪除這些資料是合理的。

5. PMM (Predictive Mean Matching)：

   我們使用 R 程式語言的 MICE(Multivariate Imputation by Chained Equations) 套件，PMM 是一種基於模型的資料插補方法，它透過建立預測模型來預測遺失值，並根據預測結果從現有的觀察值中選擇一個最接近的平均值進行填補。

   以下簡單說明 MICE 的優缺點：

   - 優點：能夠適用於多變量數據，並且可以保留數據間的相關性。處理

可以非常態分佈的數據，可以用於分類和恢復問題。

- 缺點：高維度資料的計算複雜度。由於需要對每個變數進行插值，因此隨著變數增加，計算量也會大大增加，此外，MICE 對於遺失值類型的假設比較嚴格，如果與假設不符，可能會導致插值結果不準確。

而 MICE 中 PMM 方法的優點在於能夠考慮其他變數之間的相關性。可以保持資料的分佈特性和變異性，使得插補後的資料更接近真實情況。簡單說明 PMM 的插補過程，詳見圖 8 ，現在有 A、B、C 三個變數，共有多筆資料，圖中列出前八筆，假設現在 A 變數中有兩筆資料有遺失值，則先計算所有資料在給定 B、C 變數的資料下，A 變數的期望值，接著比較這些期望值，尋找最相近的，則就用最近期望值的那筆 A 變數資料去填補遺失值，比如說第一個遺失值的 $E(A|B,C)$ 為 0.47，與其他相比最近的為 0.49，因此這筆遺失值將填補第一筆中 A 變數的資料 0.93。需要注意的是，因為 PMM 方法是用其餘變數來做預測，因此變數彼此要有相關，使用該方法的前提假設為遺失值必須屬於隨機遺失 (MAR)。

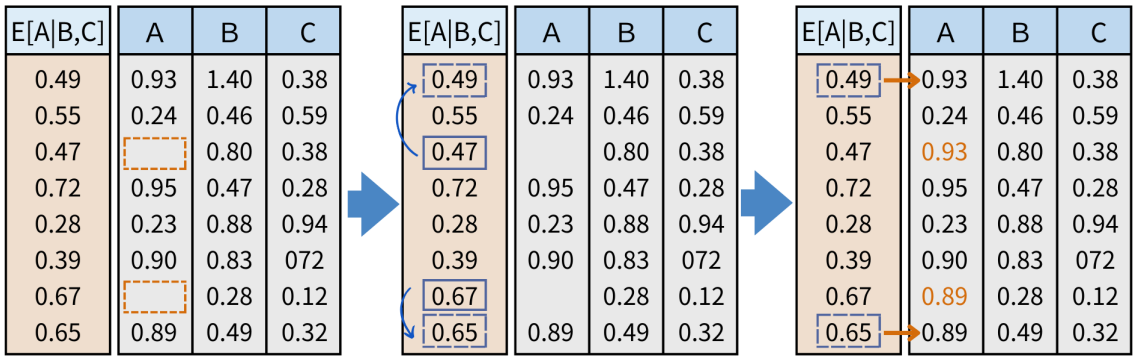


圖 8: PMM 概念

先簡單敘述插補的步驟：

-Step 1. 使用 PMM 插補五次

-Step 2. 繪畫五次插補的密度分配圖

-Step 3. 目測觀察後挑選與原始資料最像的

-Step 4. Kolmogorov-Smirnov test

-Step 5. 看統計量 D 比較五次插補

-Step 6. 確定最終選擇

(a) 第一次 PMM 插補：

- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_YEAR

AMT 系列的六個變數遺失值皆為 41376 筆，佔全部變數 13.5%。這些遺失值來自同一組資料，於前面小節所述，可知六個變數之間是息息相關的，因此遺失值屬於 MAR，服從 PMM 插補法的前提假設。

接著使用 PMM 插補法將所有遺失值填補。為了使插補更準確，我們進行了五次迭代，挑選最合適的一次作為最終插補的依據。產生的密度分配圖如圖 9 所示，其中藍色線代表原始資料的密度分配圖，紅色線依序代表插補五次的密度分配圖。從圖中觀察到，前五個變數的分佈與原始分配相似，顯示插補的效果良好；然而，最後一個變數的分佈則較為複雜，因此我們決定以 AMT_REQ_CREDIT_BUREAU_YEAR 的插補表現來決定我們選擇哪一次插補結果。

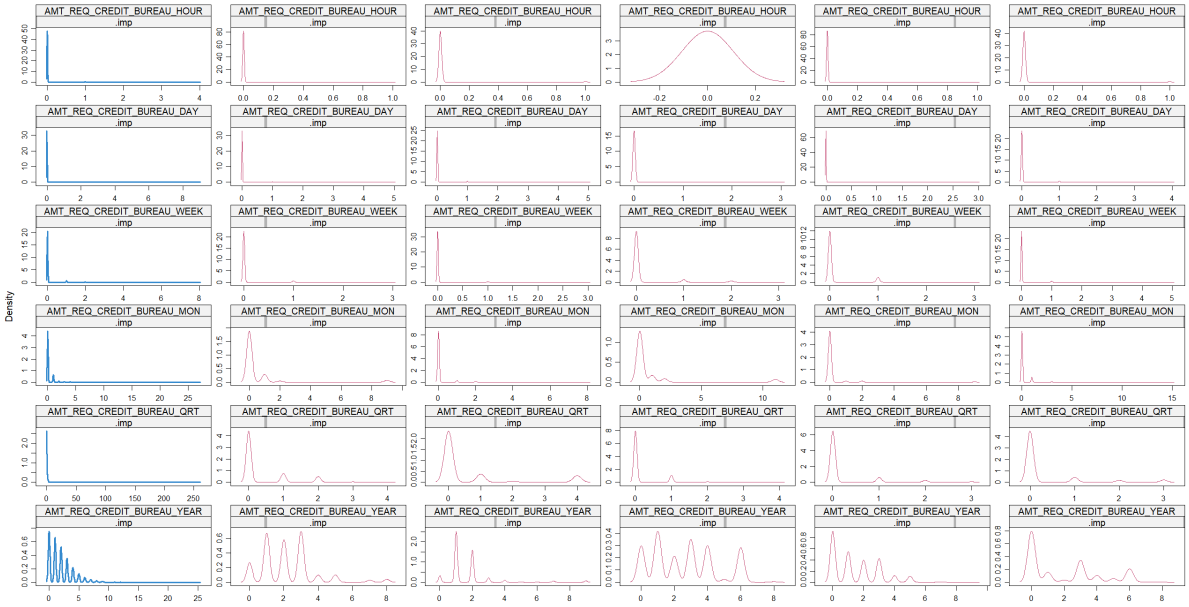


圖 9: 六個變數 PMM 插補密度分配圖

觀察圖 10 ，可以發現五次迭代的結果皆與原始資料有所差異，但第四次迭代的表現較為出色，因此我們打算以第四次迭代的結果作為我們最終插補 AMT 系列六個變數遺失值的方式。

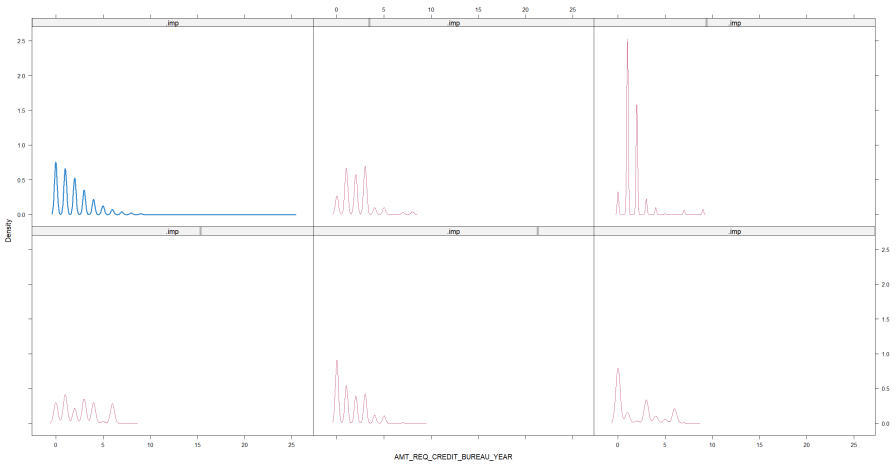


圖 10: AMT-YEAR 的 PMM 插補密度分配圖

然而，目前我們僅靠目測的方式進行選擇，為了驗證所選擇的插補結果是否合適，我們使用 Kolmogorov-Smirnov Test，觀察到 `AMT_REQ_CREDIT_BUREAU_YEAR` 原始分配就可以發現到相較於前五個變數複雜很多，預料插補後的分佈與原始分佈將有所不同，所以 p-value 皆會非常小，為了挑選五次迭代中最佳的結果，我們比較統計量 D，該統計量是基於兩個累積密度函數 (CDF) 之間的最大垂直差異（如圖 11 ），$F_{1,n}$ 和 $F_{2,m}$ 分別是第一個和第二個樣本的經驗分配函數，n, m 為其樣本數。其計算公式如下：

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

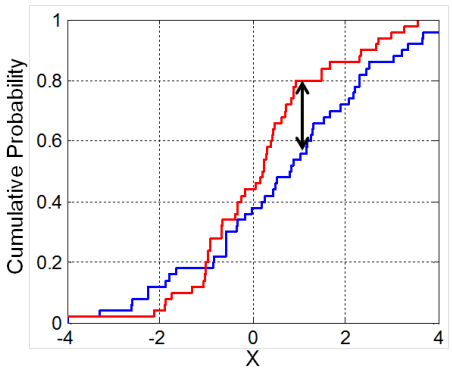


圖 11: K-S Test 統計量示意圖

如果將兩個分配的機率密度函數疊在一起，則統計量表示兩者之間差距的面積和，數值越小代表兩個分佈越相似。

H_0:The ith distribution is not significantly different from the original distribution, $i = 1, 2, 3, 4, 5$.

H_1:The ith distribution is significantly different from the original distribution, $i = 1, 2, 3, 4, 5$.

表 2: K-S Test(AMT-YEAR)

第 i 次迭代 / 統計量	1	2	3	4	5
D	0.0219	0.0275	0.0277	0.0126	0.0265

在檢定後，結果如表 2 ，確實是第四次迭代表現最好 (D 最小)，圖 12 左圖將五次迭代的 pdf 與原始分配的 pdf 放在同一張圖上，而右圖最終挑選的第四次迭代的 pdf 與原始分配的 pdf 放在同一張圖上，以便觀察。回頭看前五個變數，做完檢定後也發現我們所挑選的 p-value 趨近於 1，代表其高度拒絕 H_1 的假設，說明我們有高度證據去解釋其與原始變數極相似，因此我們決定以第四次迭代的結果作為我們最終插補 AMT 系列六個變數遺失值的方式。

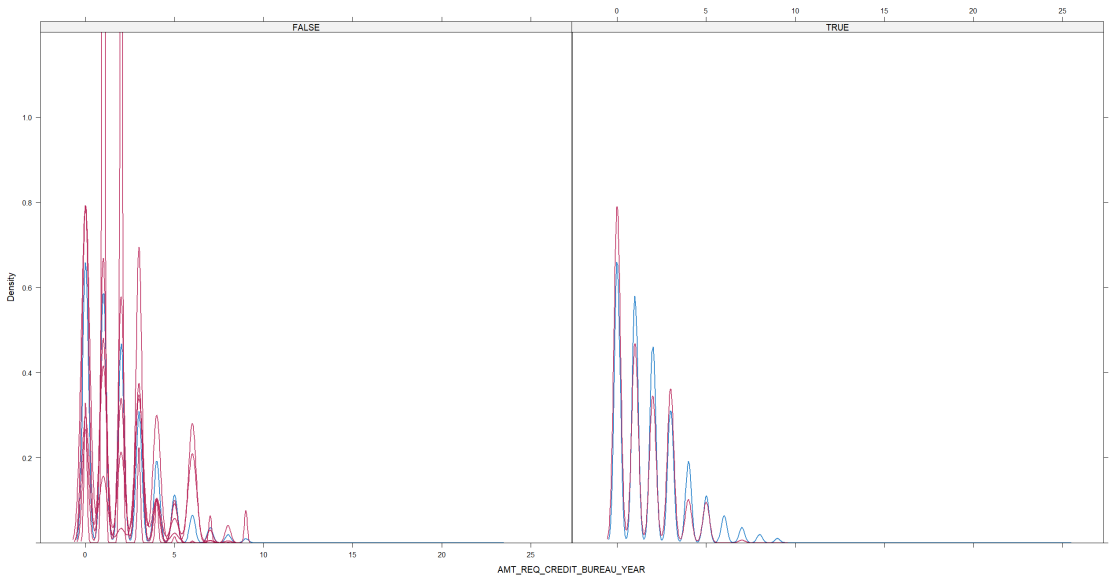


圖 12: AMT-YEAR 的 PMM 最終插補密度分配圖

(b) 第二次 PMM 插補：

因為 PMM 的方法是要使用現有資料生成一個模型後再使用這個模型去對遺失值插補，且每次使用皆須至少兩個有遺失值的變數，但目前我們僅剩餘EXT_SOURCE_3尚未插補，為了更好去預測EXT_SOURCE_3的遺失值，我們觀察相關係數圖發現所有數值型變數與其他變數的相關性皆不高，如附錄圖 32 ，但在當中做比較的話，YEAR_BIRTH與其相關性最高，因此將YEAR_BIRTH放入我們的插補模型中，又因YEAR_BIRTH原始資料是完整的，因此我們先將YEAR_BIRTH隨機產生 10% 的遺失值以順利進行 PMM 插補。其於步驟與第一次插補的方式一樣。其中因為兩者是有相關性的，表示其遺失值屬於 MAR，服從 PMM 插補法的前提假設。

- EXT_SOURCE_3 (60771 筆，佔全部變數 19.8%)
- YEAR_BIRTH(自行生成 10% 遺失值)

接著使用 PMM 插補法將所有遺失值填補。為了使插補更準確，我們進行了五次迭代，挑選最合適的一次作為最終插補的依據。產生出的密度分配圖如圖 13 所示，其中藍色線代表原始資料的密度分配圖，紅色線依序代表插補五次的密度分配圖。因為 YEAR_BIRTH 有完整的資料，在生成 10% 遺失值後也低於 EXT_SOURCE_3 的 19.8%，所以插補的表現前者會表現較好，因為其有較多的資訊去生成新資料，因此我們決定以後者的插補表現來決定我們選擇哪一次插補結果。

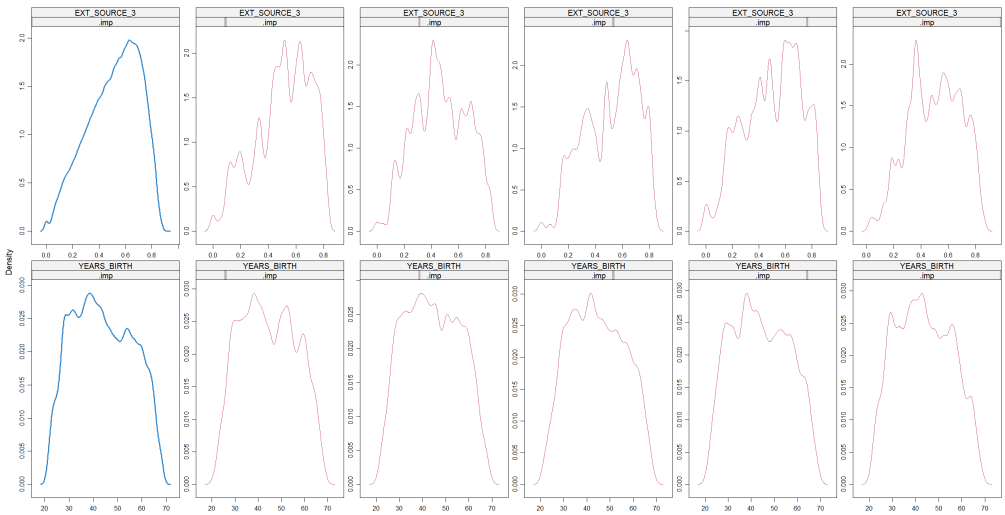


圖 13: 兩個變數 PMM 插補密度分配圖

觀察圖 14 ，發現五次迭代的結果皆與原始資料有所差異，但第四次迭代的表現較為出色，因此我們打算以第四次迭代的結果作為我們最終插補這兩個變數遺失值的方式。

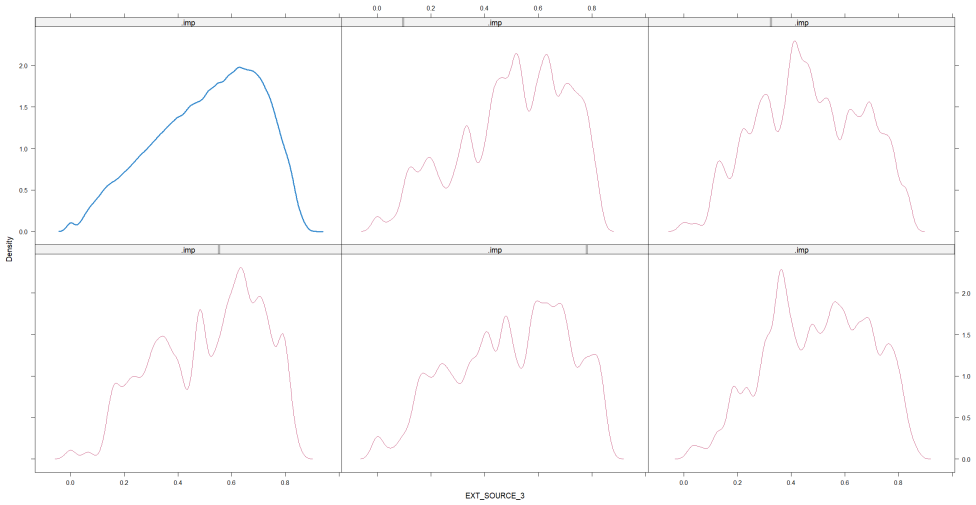


圖 14: EXT-3 的 PMM 插補密度分配圖

然而，目前我們僅是靠目測的方式進行選擇，為了驗證所選擇的插補結果是否合適，我們一樣使用 K-S Test：

H_0:The ith distribution is not significantly different from the original distribution,$i = 1, 2, 3, 4, 5$.

H_1:The ith distribution is significantly different from the original distribution,$i = 1, 2, 3, 4, 5$.

表 3: K-S Test(EXT-3)

第 i 次迭代 / 統計量	1	2	3	4	5
D	0.0030	0.0020	0.0028	0.0014	0.0025

在檢定後，結果如表 3 ，確實是第四次迭代表現最好 (D 最小)，圖 15 左圖將五次迭代的 pdf 與原始分配的 pdf 放在同一張圖上，而右圖最終挑選的第四次迭代的 pdf 與原始分配的 pdf 放在同一張圖上，以便觀察。因此我們決定以第四次迭代的結果作為我們最終插補 AMT 系列六個變數遺失值的方式。

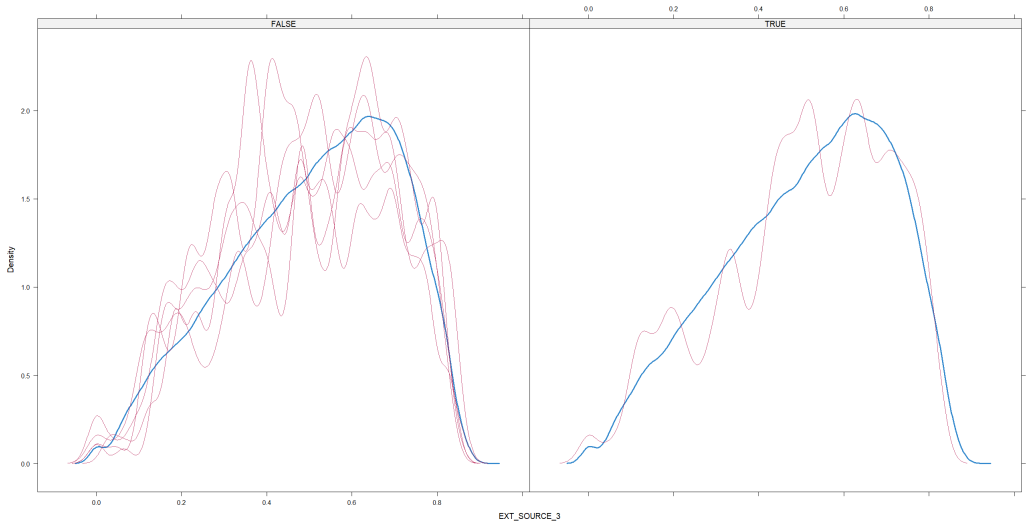


圖 15: EXT3 的 PMM 最終插補密度分配圖

3.5 不平衡資料處理

在資料介紹中，我們提到了當資料集中的目標變數比例出現懸殊時，需要進行處理。這種情況下，常見的處理方式有兩種：Undersampling 和 Oversampling。簡而言之，Undersampling 是刪除多數類別的資料以平衡比例，而 Oversampling 則是增加少數類別的資料來實現平衡。經過一系列嘗試後，我們發現無論採取何種特殊方法，結果都差不多，因此我們將採取最簡單的 Random Undersampling 及 Random Oversampling 將資料中有違約與無違約的資料筆數調整至 1:1。抽樣執行完成後我們將結果分成三個資料集去比較，分別為原始資料集與經過隨機欠抽樣、隨機過抽樣的資料集。接著使用 K-fold Validation 方法將三個資料集分別切割成四層，再分別利用羅吉斯迴歸 (Logistic Regression)、決策樹 (Decision Tree) 以及隨機森林 (Random Forest) 三個方法各執行四次 (C_3^4)，並產生 AUC 值以進行比較。

其中需要注意的是，不平衡資料在使用 K-fold 切割時有可能會發生某一層的資料沒有被分配到某個二元分類變數的其中一個值，若確實有該情況發生則無法進行交叉驗證。為了避免此狀況，我們在進行交叉驗證前，會先將一些變數去除來確保後續的步驟能夠正常執行。

表 4 為其交叉驗證結果，我們分別將隨機欠抽樣與隨機過抽樣後的資料集所產生的三個 AUC 值來與原始資料集的產生的三個 AUC 值比較，發現兩種抽樣處理方法後的資料集之 AUC 值表現都與原始資料集之 AUC 值來的好。但可以看到原始資料使用決策樹方法的表現極差，AUC 值為 0.5，其預測正確機率與丟

銅板一樣，而抽樣後的資料表現雖然有比較好，但也不及 0.7，因此判斷此資料並不適合使用決策樹的方法。因此接下來我們比較羅吉斯迴歸及隨機森林的方法，發現隨機欠抽樣與隨機過抽樣後的資料集所產生的 AUC 值互相比較的話，隨機過抽樣的資料集表現的較好，因此後續依序做的特徵選取以及建立模型皆是採用隨機過抽樣後產生的資料集。

表 4: K-fold Validation 的 AUC

模型方法 / 資料集	Logistic Regression	Decision Tree	Random Forest
Raw Data	0.73118	0.5	0.708897
Undersampling Data	0.730358	0.645275	0.7068887
Oversampling Data	0.731453	0.646103	0.7112601

因為在交叉驗證前有排除幾個變數，我們發現其中被排除的變數皆為類別型變數，為了探討被排除後的影響，我們想查看這幾個變數與目標變數 (TARGET) 之間的相關性，使用的是 Mutual Information，以下簡單介紹一下：

交互信息 (Mutual Information)[1] 是一種衡量兩個隨機變數之間相依性的指標。它衡量的是一個隨機變數中包含的關於另一個隨機變數的信息量。當兩個變數之間的交互信息越大，表示它們之間的相依性越強。

具體來說，如果我們有兩個隨機變數 X 和 Y，它們的交互信息可以表示為 $I(X;Y)$，其計算公式如下：

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

其中，$p(x,y)$ 是隨機變數 X 和 Y 同時取值 x 和 y 的機率，$p(x)$ 和 $p(y)$ 分別是 X 和 Y 的邊際機率。這個公式可以理解為，交互信息衡量的是 X 和 Y 同時取值的聯合機率分佈與它們各自獨立取值的機率分佈之間的差異。

而對於我們目前的三筆資料，觀察到這些被去除的變數與目標變數 (Target) 的 Mutual Information 皆很低（詳見附錄表 10 ），因此將其事先去除是可行的，對模型不會有太大的影響。

[1]在計算上，Entropy（熵）是關於單個隨機變數的，而 Mutual Information（交互信息）則是關於兩個隨機變數的。在應用上，Entropy（熵）通常用於分類或分群任務中的不確定性衡量，而 Mutual Information（交互信息）常用於特徵選擇或特徵相關性分析中。

4 特徵選取

以下是我們的變數篩選方式，分別將連續型及類別型變數分開討論，前者採用 Point-Biserial Correlation，後者採用 Mutual Information 的方法，兩種類型的變數各選擇對目標變數重要性的前五名。

- 類別型變數 (TARGET) VS 數值型變數：使用 Point-Biserial Correlation。

  Point-Biserial Correlation 是一種用於衡量一個二元變數和一個連續變數之間關係的統計方法。其計算過程包括以下步驟：

-Step 1. 計算二元變量的平均值（即佔比），以及連續變量的平均值。

-Step 2. 計算二元變量的標準差，以及連續變量的標準差。

-Step 3. 計算兩個變量的共變異數。

-Step 4. 使用以下公式計算 Point-Biserial Correlation：

$$r_{pb} = \frac{M_1 - M_0}{s_y} \sqrt{\frac{N_0 N_1}{N(N-1)}}$$

其中：

- r_{pb} 是 Point-Biserial Correlation 係數。
- M_1 和 M_0 是連續變數在二元變數分組中的平均值。
- s_y 是連續變數的標準差。
- N_1 和 N_0 是分別屬於二元變數的兩個分組的樣本大小。

-Step 5. 通常，Point-Biserial Correlation 的值介於-1 和 1 之間。值越接近 1 或-1，表示二元變數和連續變數之間的關係越強。如果 Point-Biserial Correlation 接近 0，則表示兩者之間幾乎沒有相關性。

就我們目前的資料來說，首先先將目前有的變數中數值型的變數挑選出來做 Point-Biserial Correlation，表 5 為其相關性前十名的結果，首先先看 Pearson's Product-Moment Correlation Test，運用相關性檢定，確認輸出的相關性值是否可信。數值型變數完整結果詳見附錄表 **??** 。

H_0:True correlation is equal to 0.

H_1:True correlation is not equal to 0.

可以發現前五名的變數，p-value 皆趨於 0，表示我們有足夠證據去說明其與目標變數確實是有相關的，也就代表可以直接看相關性的值去比較數值型變數對目標變數的重要性。

表 5: 類別型變數 (TARGET) VS 數值型變數

數值型變數	p-value	Correlation	名次
EXT-SOURCE-2	0	-0.268909	1
EXT-SOURCE-3	0	-0.242487	2
YEARS-BIRTH	0	-0.137025	3
YEARS-LAST-PHONE-CHANGE	0	-0.10221	4
YEARS-ID-PUBLISH	0	-0.093831	5
YEARS-REGISTRATION	0	-0.079577	6
AMT-GOODS-PRICE	0	-0.07716	7
REGION-POPULATION-RELATIVE	0	-0.07172	8
AMT-CREDIT	0	-0.058562	9
DEF-30-CNT-SOCIAL-CIRCLE	0	0.055519	10

- 類別型變數 (TARGET) VS 類別型變數：使用 Mutual Information。

我們挑選與目標變數交互信息前五名的類別型變數，將這幾個變數留下並做後續建立模型的步驟，表 6 列出前十名的數值。

表 6: 類別型變數 (TARGET) VS 類別型變數

類別型變數	$I(X;Y)$	名次
OCCUPATION-TYPE	0.010219	1
ORGANIZATION-TYPE	0.00829	2
NAME-INCOME-TYPE	0.007373	3
REGION-RATING-CLIENT-W-CITY	0.006174	4
NAME-EDUCATION-TYPE	0.006086	5
REGION-RATING-CLIENT	0.005799	6
CODE-GENDER	0.004725	7
FLAG-EMP-PHONE	0.003968	8
REG-CITY-NOT-WORK-CITY	0.003497	9
FLAG-DOCUMENT-3	0.002722	10

根據以上特徵選取，我們最終挑選的十個變數如表 7。

表 7: 最終選取變數

數值型變數	類別型變數
EXT-SOURCE-2 1	OCCUPATION-TYPE
EXT-SOURCE-3	ORGANIZATION-TYPE
YEARS-BIRTH	NAME-INCOME-TYPE
YEARS-LAST-PHONE-CHANGE	REGION-RATING-CLIENT-W-CITY
YEARS-ID-PUBLISH	NAME-EDUCATION-TYPE

5 分析方法

在前面小節中我們有討論羅吉斯迴歸、決策樹、隨即森林對資料訓練的 AUC 值，其中羅吉斯回歸表現最好，因此後續的部分使用羅吉斯迴歸。在這邊第一步是將訓練資料使用 k-fold 切分成 4 層，放入羅吉斯迴歸模型中。接著使用兩種不同的方法挑選切點。

- 觀察每一層 K-fold AUC 值，第四層 AUC 值為 0.738，其餘三層皆為 0.734，因此選擇第四層中 F1 值表現最佳的點為 0.282，如圖 16 ，其對應的切點為 0.14，所以在此方法中使用 0.14。
- 先計算所有可能切點在每層 K-fold 中的 AUC 值，對於每個不同的切點計算四層的平均 AUC 值，選擇使平均 AUC 值最大的切點，因此挑選切點為 0.13，F1 值為 0.283。

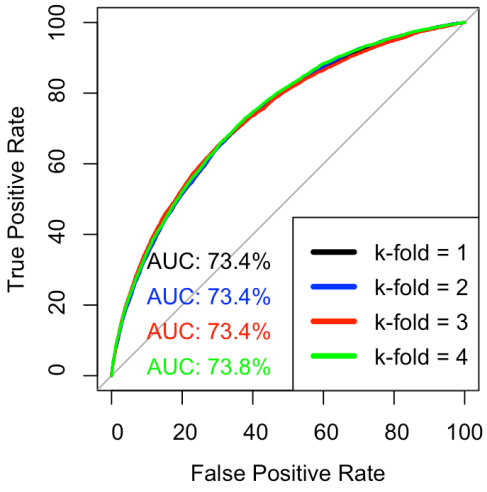


圖 16: K-fold 四層 AUC 值

先介紹稍後會參考的指標：

1. Sensitivity 或 Recall 值：有違約樣本在所有正確預測樣本中的比例。
2. Precision 或 PPV(Positive Predictive Value)：有違約的樣本中被正確預測為有違約的精確度。
3. F1 值：其為上面兩者的調和平均數，對於我們的不平衡資料而言，他提供一個比 Accuracy 更適合的參考指標。
4. Accuracy：模型預測正確數量所佔整體的比例。

此專案較看重的是指標為 PPV 與 F1 值，其中若選擇數值較高的 PPV 值來讓有違約的樣本被正確預測的機率提升的話，反之會造成 Recall 值降低，可能無法正確辨識到有違約的樣本，同時也可能降低 F1 值，因為其考慮的是兩者之間的平衡，數值表現如圖 8 。對於現實中的應用來說，可以選擇使用接近 1 的 PPV 值來讓有違約的樣本被正確預測的機率接近 1，但會花很高的成本來去看檢查每個樣本的違約與否。因此我們在後續的網頁中主要會使用 0.13 作為我們選取的切點，雖然準確率較使用 0.14 作為切點低一些，但我們希望有違約的樣本被正確預測的機率高一些。

表 8: 各種指標表現

Threshold	Sensitivity/Recall	Precision/PPV	F1	Accuracy
0.13	0.237	0.47	0.315	0.773
0.14	0.253	0.43	0.319	0.795

6 會議紀錄

- 口頭報告 QA 與建議：
  1. 第一次：
     - Q：會採取 Oversampling 來處理不平衡資料嗎？

       A：尚未決定，會到後續做測試後在選擇要使用何種抽樣去處理不平衡資料。
     - 建議：三個遺失值類型中非隨機遺失較多，需著重看資料背景的再決定如何處理遺失值。

- Q：變數合併的理由 FLAG_DOCUMENT、AMT_REQ_CREDIT_ ?

  A：原先是想說前者不知道文件內容為何，解釋上會很難表示，所以希望合併成一個新變數並將原本的二十個變數刪除，但考慮老師提供的建議，實務上或許是因為涉及商業機密才不得提供資料內容，因此我們決定新增一個變數後並不刪除原先變數。後者也會保留原先變數。

- Q：為什麼取前十名變數？

  A：因為目前變數有近兩百個，為了排除不重要的變數及遺失值過多的變數，我們會先挑選前十名重要的變數放入模型中，再去看其表現效果，若很糟的話會是情況增加變數。

- Q：有訓練模型後可以會拿來與 kaggle 比對嗎？

  A：不會，因為 Kaggle 上的大家的做法都沒有考慮到不平衡資料需做處理，所以我們並不會與其做比對。

2. 第二次：

   - 表現很好沒有什麼問題。

3. 第三次：

   - 表現很好沒有什麼問題。

• 小組討論紀錄：

1. 3/5：

   - 確認職位分工: Project Manager(楊廷紳、林貫原)、Data Scientest(易祐辰、留筠雅)、System Developer(許政揚、周昱宏)
   - 討論研究主題與目的：考慮到學習目標:Missing values、Many categorial variables with many levels、Classification、Clustering

     暫定：預測客戶有無信用卡違約 (詐欺)(變數: Target = 1、0，1 為違約)、影響信用卡違約的主要變數有哪些?

     目的: 讓銀行決定要不要批准客戶的借貸

– 弄清楚資料檔的使用目的：主要使用：creditcard_train、creditcard_test。變數解釋：columns_description。未知：previous_application、creditcard_test_true的使用目的

2. 3/12：

– 資料說明：

(a) 刪除 45 個變數，遺失值超過 32%(剩餘一個標準差界線的資料數，如果進行插補) 資料的變數。

(b) 是否提供變數FLAG_DOCUMENT合併成提供幾件文件、遺失值設為 0 未提供)。

– 觀察結果: 待處理完資料。

– 預期達成的目標：

(a) 運用 PCA 分群出重要的變數以判斷是否違約。

(b) 建立分群前與分群後的模型比較。

(c) Rshiny 網頁 demo(將已完成的分析展示)。

3. 3/19：

– 變數選擇：確定將簽署文件與否合併為簽署文件多寡、討論類別型變數的做法。

– 遺失值插補方法。

– 第一次專案報告前的進度規劃確認：

(a) 3/26 確認多種插補方法的使用時機、優點、缺點，決定該資料適合使用何種方法。

(b) 4/2 將遺失值全數插捕完成。

(c) 4/7 完成變數介紹、專案 ppt 並模擬報告。

4. 3/26：

– 專案背景說明、專案執行計畫、甘特圖、組員分工由 PM 撰寫。

- 資料說明與觀察結果由 DS 撰寫。
- 預期達成目標由 SD 撰寫。
- 資料整理：
  (a) 確認是否有重複資料。
  (b) 判斷各種變數是否有不合理資料。
  (c) 轉換數值與類別變數。

5. 4/7：第一次專案報告前模擬報告、針對投影片內容做修正。
6. 4/13：
   - 第一次專題報告問題檢討。
   - 後續資料處理流程。
   - 變數討論的結論：`FLAG_DOCUMENT`原本的 20 個變數要留下並額外加一個加總的變數、AMT 六個變數皆留下並新增一個調查一整年的次數
7. 4/16：
   - 更改工作分配：Project Manager(許政揚、林貫原)、Data Scientest(留筠雅)、System Developer(楊廷紳、周昱宏)、Technical Support(易祐辰)。
   - PM 介紹幾種不平衡資料處理方法。
   - DS 介紹插補不同變數的方法。
   - TS 秀視覺化圖形。
   - SD 提一下網頁可以放的東西。
   - 討論後續分析問題：
     (a) 討論離群值呈現方式。
     (b) 討論 EDA 要呈現的東西。
     (c) 討論資料降維 (PCA, 相關係數)。

7　網頁使用說明書

本儀表板旨在以客戶基本資料與交易資料為基礎，呈現信用卡違約風險因素分析與預測。圖 17 為本儀表板首頁，以下則為主要功能介紹：

1. 資料視覺化：

   提供多種資料相關統計圖表，便於使用者了解數據分佈及趨勢，且可以根據不同的參數進行過濾與更新。

2. 違約風險預測：

   應用模型（羅吉斯迴歸）進行信用卡違約風險預測。並提供評估指標。

3. 個案分析：

   提供單個客戶的詳細分析報告，包括違約風險評估、影響因素等。

圖 17: 儀表板首頁

7.1　數據摘要

此功能將針對類別型變數與連續型變數提供資料清洗後的統計圖表以及統計數據摘要。

7.1.1　類別型變數

圖 18 為類別型變數的統計圖表顯示頁面。由於本專案的主要反應變數是客戶是否有信用卡違約的狀況，頁面中的左側功能欄位提供了客戶群體的選擇如下，客戶群體選擇示範如圖 20。

(a) 有：僅顯示有違約之客戶之資料

(b) 無：僅顯示無違約之客戶之資料

(c) 全體：顯示所有客戶之資料

接著再選擇想要觀察的變數名稱，頁面即會顯示指定客戶群體在該變數下的類別變數長條圖。如需要放大畫面顯示，可以透過拖移游標選取欲放大檢視的部分圖表，頁面即會顯示放大後的長條圖，如圖 19。

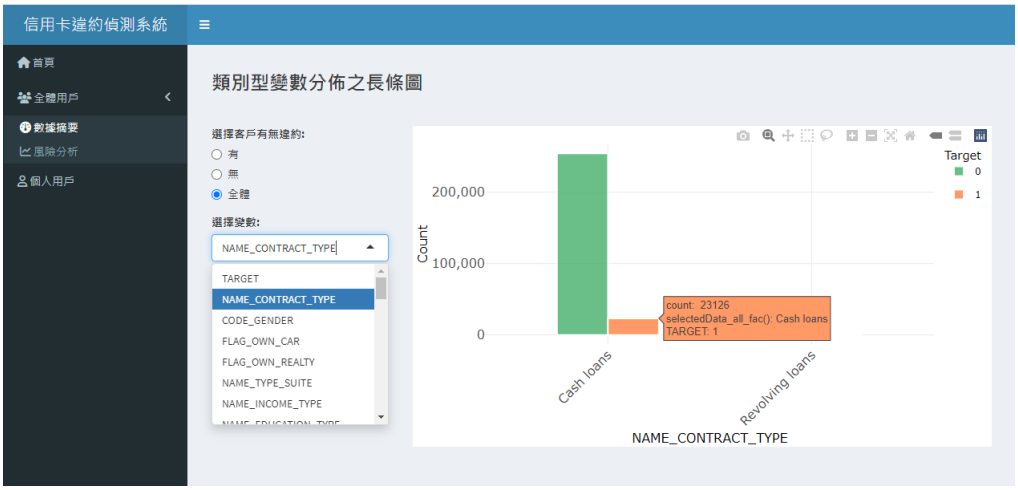


圖 18: 類別型變數介面

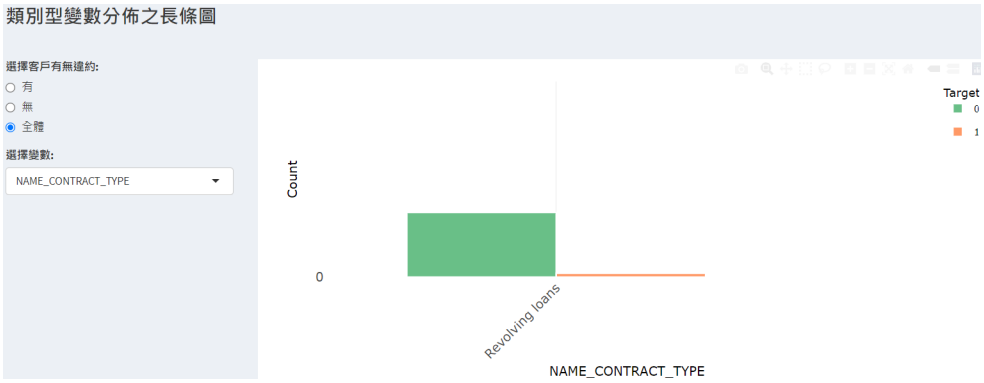
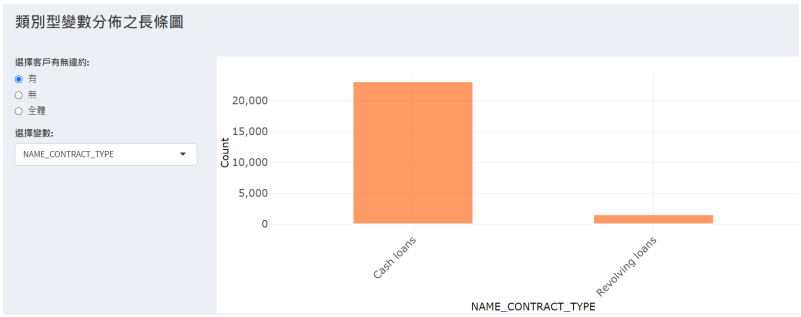
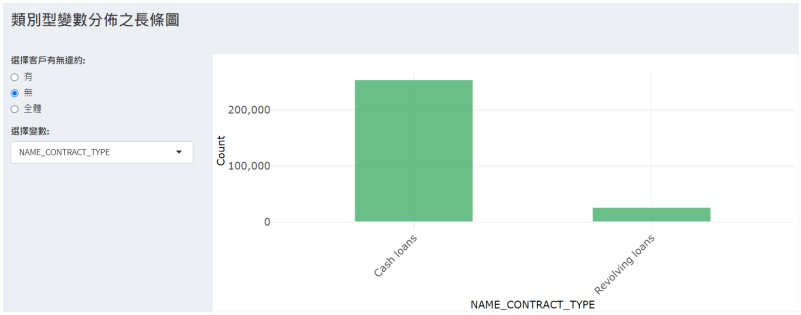


圖 19: 長條圖放大檢視示範

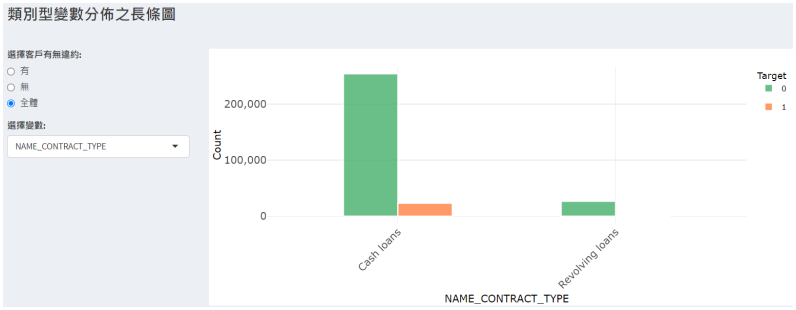
使用者如欲下載長條圖之圖檔，可點畫面右上角的下載按鈕下載圖檔，如圖 21 中紅色方框所示的下載圖檔按鈕，圖 22 則表示使用者可以選擇將欲下載的圖檔存放於自行選擇的資料中。



(a) 有違約客戶



(b) 無違約客戶



(c) 全體客戶

圖 20: 類別型變數長條圖功能介紹

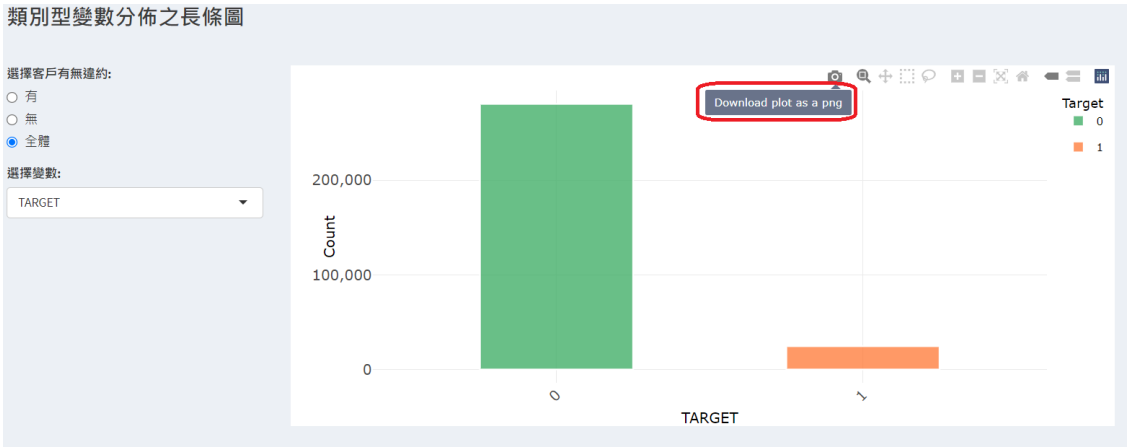


圖 21: 圖檔下載按鈕

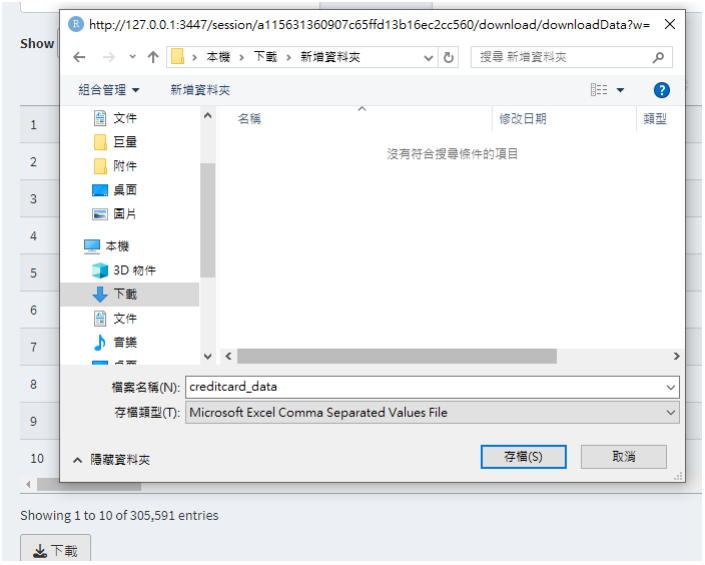


圖 22: 圖檔下載介面

7.1.2 連續型變數

圖 23 為連續型變數的統計圖表與摘要顯示頁面。頁面中的左側功能欄位同樣提供了客戶群體的選擇如下，客戶群體選擇示範如圖 24。而連續型變數在統計圖表的部分顯示的為直方圖，同時敍述統計量將會顯示在直方圖下方，將會根據指定的客戶條件顯示相對應的資訊。

(a) 有：僅顯示有違約之客戶之資料

(b) 無：僅顯示無違約之客戶之資料

(c) 全體：顯示所有客戶之資料

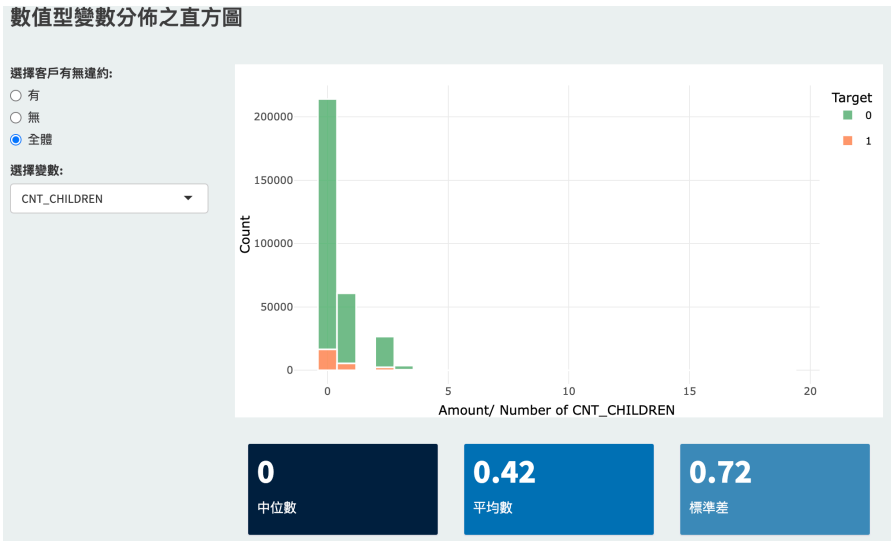
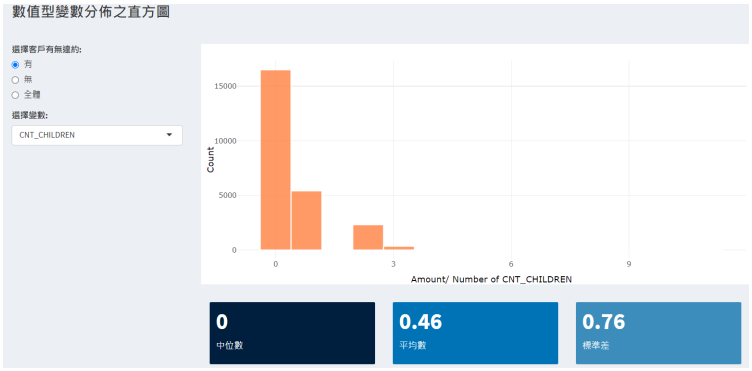
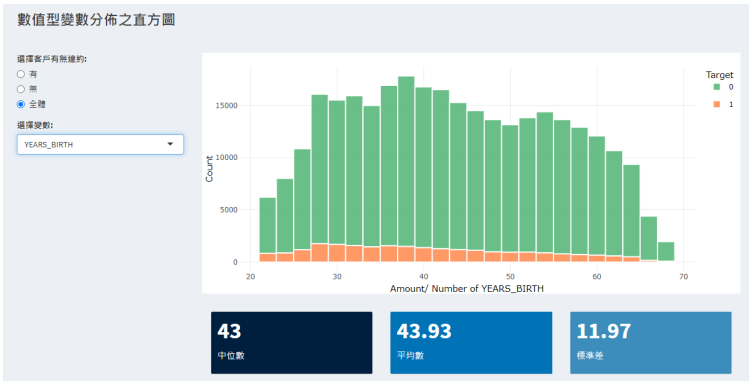


圖 23: 連續型變數介面



(a) 有違約客戶

(b) 無違約客戶



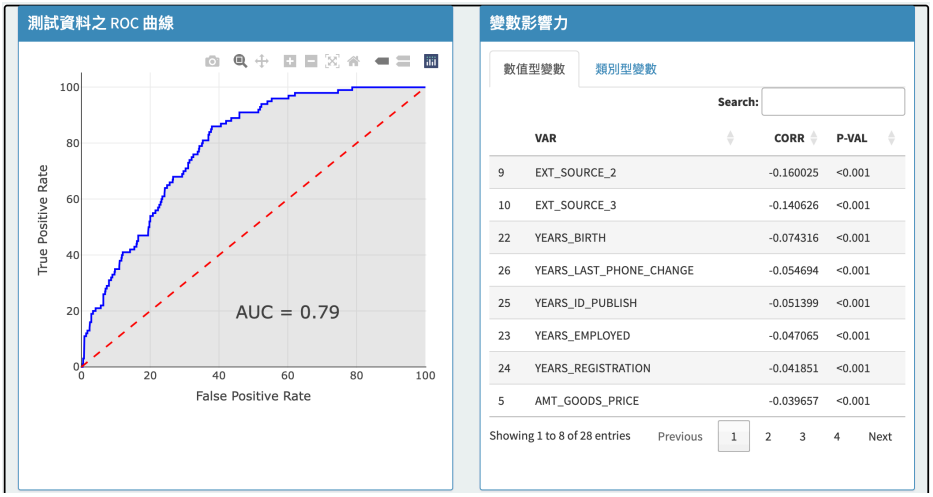
(c) 全體客戶

圖 24: 連續型變數直方圖功能介紹

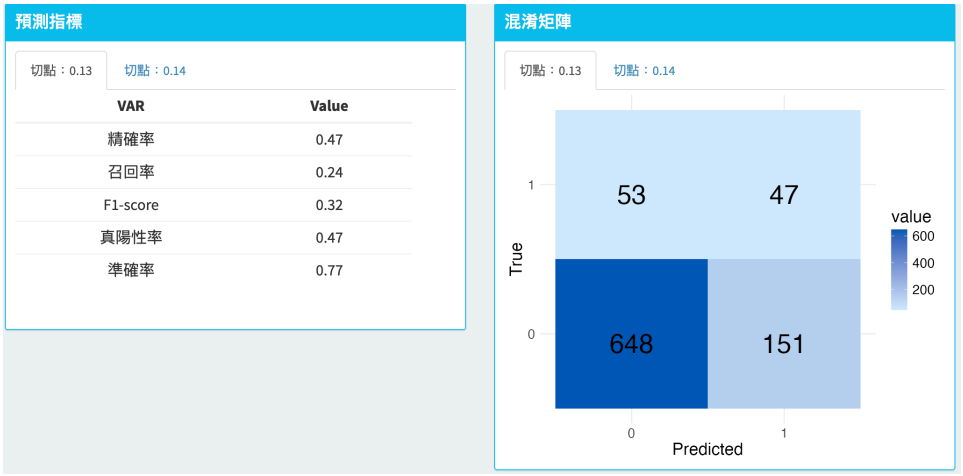
7.2 風險分析

此功能提供了本次專案的整體分析結果，圖 25 將顯示以下四個不同資訊：

- 測試資料的 ROC 曲線。
- 變數影響力。
- 預測指標。
- 混淆矩陣。



(a) ROC 曲線及變數影響力



(b) 預測指標及混淆矩陣

圖 25: 風險分析介面

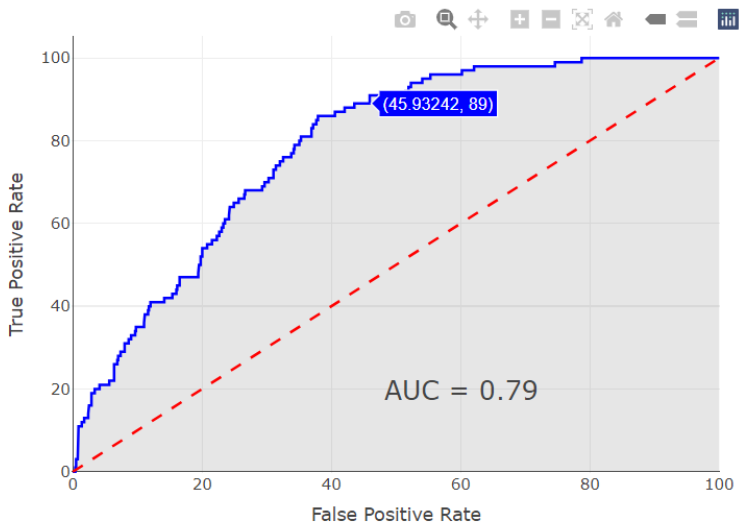


圖 26: 測試資料的 ROC 曲線

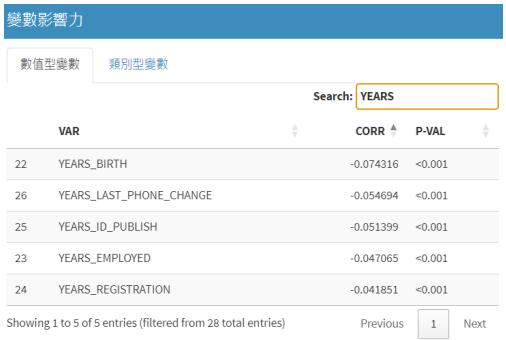
7.2.1 測試資料的 ROC 曲線

圖 26 顯示的本專案測試資料的 ROC 曲線。其展示了模型在指定閾值下的性能。通過繪製真陽性率對假陽性率的曲線，提供使用者直觀地評估模型的整體分類能力。因此在此功能中，我們提供使用者能夠將游標移動到藍色的 ROC 曲線上，畫面即會顯示對應的座標。此外，本專案模型的最終 AUC 指標之值為 0.79。

7.2.2 變數影響力

此功能提供了連續型變數和類別型變數依照影響力大小的絕對值排序。對於連續型資料而言，會再計算每個變數的相關係數並顯示其 p 值，如圖 27 (a) 所示。此功能中同事提供使用者在右上角的搜尋欄位中輸入欲觀察的變數名稱，以便觀察特定變數的變數影響力，如圖 27 (b) 所示。

(a) 類別型變數影響力



(b) 搜尋功能

圖 27: 類別型變數影響力介面

7.2.3 預測指標

此功能提供了本次專案所建立模型的精確率、召回率、F1-score、真陽性率及準確率，並將這些指標整理成一個表格，如圖 28 所示。預測指標包括準確率、精確率、召回率和 F1 分數等，提供使用者更全面的衡量指標。

7.2.4 混淆矩陣

此功能提供了本次專案所建立模型的混淆矩陣，如圖 29 所示。混淆矩陣是一個表格，提供了模型的分類結果，包括真陽性（TP）、假陽性（FP）、真陰性（TN）和假陰性（FN）的數量。通過分析混淆矩陣，使用者可以直觀地看到模型在各

個類別上的預測錯誤情況，具體了解模型在哪些類別上有較好的性能，哪些類別上存在問題。其中 0.47 的部分即是本專案中有違約的客戶被正確預測成有違約的機率。

圖 28: 預測指標表格

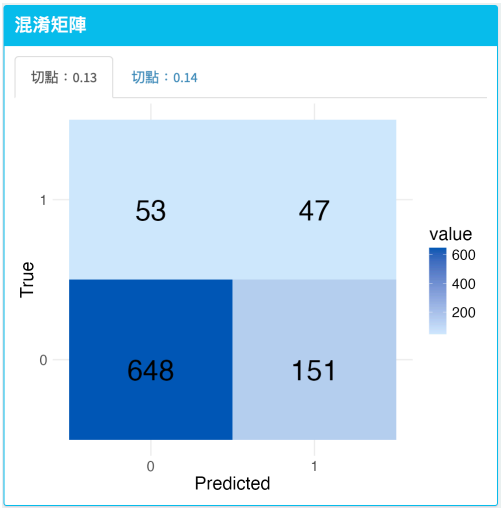
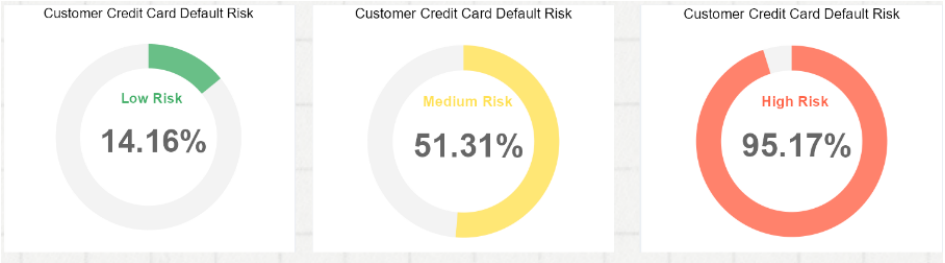


圖 29: 混淆矩陣

7.3 個案分析

此功能提供使用者預測個人的違約風險。使用者可以輸入相關資料並按下提交，如圖 30a 所示。系統就會自動利用本專案的模型計算客戶違約機率。接著將風險分為低、中、高，分別以綠色、黃色和紅色的甜甜圈圖表示，如圖 30b 所示。

(a) 輸入資料



(b) 違約機率及風險高低

圖 30: 個案分析介面

8 附錄

以下為所有變數的介紹，* 代表的是經過資料處理過後所留下的變數，若無特殊備註，詢問是否的變數，1 皆代表是，0 皆代表否。

1. *SK_ID_CURR：ID。
2. *TARGET：目標變數（1：至少有一次延遲付款；0：沒有延遲付款過）。
3. *NAME_CONTRACT_TYPE：現金型信貸還是循環型信貸[2]。
4. *CODE_GENDER：性別。
5. *FLAG_OWN_CAR：客戶是否擁有汽車。
6. *FLAG_OWN_REALTY：客戶是否擁有土地或房屋。
7. *CNT_CHILDREN：客戶的小孩數量。
8. *AMT_INCOME_TOTAL：客戶收入。
9. *AMT_CREDIT：客戶信用卡額度。
10. *AMT_ANNUITY：貸款年金。
11. *AMT_GOODS_PRICE：欲貸款購買商品之價格。
12. *NAME_TYPE_SUITE：客戶在申請貸款時的陪同人員。
13. *NAME_INCOME_TYPE：工作類型。
14. *NAME_EDUCATION_TYPE：教育程度。
15. *NAME_FAMILY_STATUS：家庭情況。
16. *NAME_HOUSING_TYPE：居住狀況。
17. *REGION_POPULATION_RELATIVE：客戶所居住地區的人口規模標準化值（值越大人口越多）。
18. *DAYS_BIRTH：客戶年齡（以天計算）。
19. *DAYS_EMPLOYED：申請前該客戶開始現職工作天數。

[2]循環型信貸又稱「理財型貸款」，允許借款人在指定額度內隨借隨還，只需支付動用額度的利息。此類貸款無需綁約、抵押品或保證人，且提前清償無違約金，讓資金使用更靈活。

20. *DAYS_REGISTRATION：客戶在申請前多少天更改申請資料。
21. *DAYS_ID_PUBLISH：客戶在申請前多少天變動身分證明文件。
22. OWN_CAR_AGE：客戶車齡。
23. *FLAG_MOBIL：客戶是否提供手機號碼。
24. *FLAG_EMP_PHONE：客戶是否提供工作用電話號碼。
25. *FLAG_WORK_PHONE：客戶是否提供公司電話。
26. *FLAG_CONT_MOBILE：手機是否接通。
27. *FLAG_PHONE：客戶是否提供家用電話。
28. *FLAG_EMAIL：客戶是否提供電子郵件。
29. *OCCUPATION_TYPE：客戶職業。
30. *CNT_FAM_MEMBERS：客戶家庭成員數量。
31. *REGION_RATING_CLIENT：對客戶所在地區評分 (1、2、3)。
32. *REGION_RATING_CLIENT_W_CITY：對客戶所在城市評分 (1、2、3)。
33. *WEEKDAY_APPR_PROCESS_START：客戶在星期幾的時候申請。
34. *HOUR_APPR_PROCESS_START：客戶大約在什麼時間申請。
35. *REG_REGION_NOT_LIVE_REGION：根據地區評分，客戶的永久地址與聯絡地址是否相同（1：不同；0：相同）。
36. *REG_REGION_NOT_WORK_REGION：根據地區評分，客戶的永久地址與工作地址是否相同（1：不同；0：相同）。
37. *LIVE_REGION_NOT_WORK_REGION：根據地區評分，客戶的聯絡地址與工作地址是否相同（1：不同；0：相同）。
38. *REG_CITY_NOT_LIVE_CITY：根據城市評分，客戶的永久地址與聯絡地址是否相同（1：不同；0：相同）。
39. *REG_CITY_NOT_WORK_CITY：根據城市評分，客戶的永久地址與工作地址是否相同（1：不同；0：相同）。

40. *LIVE_CITY_NOT_WORK_CITY：根據城市評分，客戶的聯絡地址與工作地址是否相同（1：不同；0：相同）。
41. *ORGANIZATION_TYPE：客戶公司類型。
42. EXT_SOURCE_1：外部數據來源 1 的正規化分數。
43. *EXT_SOURCE_2：外部數據來源 2 的正規化分數。
44. *EXT_SOURCE_3：外部數據來源 3 的正規化分數。
45. APARTMENTS_AVG：客戶所居的建築物經過正規化後的平均面積。
46. BASEMENTAREA_AVG：客戶所居建築物地下室經過正規化後的平均面積。
47. YEARS_BEGINEXPLUATATION_AVG：客戶所居建築物經正規化後平均施工年數。
48. YEARS_BUILD_AVG：客戶所居建築物正規化後的平均建築年齡。
49. COMMONAREA_AVG：客戶所居建築物正規化後平均公共區域面積。
50. ELEVATORS_AVG：客戶所居建築物正規化後平均電梯數量。
51. ENTRANCES_AVG：客戶所居建築物正規化正規化後平均出入口數量。
52. FLOORSMAX_AVG：客戶所居建築物正規化後平均最高樓層。
53. FLOORSMIN_AVG：客戶所居建築物正規化後平均最低樓層。
54. LANDAREA_AVG：客戶所居建築物正規化後平均土地面積。
55. LIVINGAPARTMENTS_AVG：客戶居住用建築物正規化後平均數。
56. LIVINGAREA_AVG：客戶居住用建築物正規化後居住平均區域面積。
57. NONLIVINGAPARTMENTS_AVG：客戶非居住用建築物經正規化後平均數。
58. NONLIVINGAREA_AVG：客戶非居住用建築物正規化後居住平均區域面積。
59. APARTMENTS_MODE：客戶所居建築物正規化後的面積大小的眾數。
60. BASEMENTAREA_MODE：客戶所居建築物地下室正規化後的面積眾數。

61. YEARS_BEGINEXPLUATATION_MODE：客戶所居建築物正規化後施工年數眾數。
62. YEARS_BUILD_MODE：客戶所居建築物正規化後的建築年齡眾數。
63. COMMONAREA_MODE：客戶所居建築物正規化後公共區域面積眾數。
64. ELEVATORS_MODE：客戶所居建築物正規化後電梯數量眾數。
65. ENTRANCES_MODE：客戶所居建築物正規化正規化後出入口數量眾數。
66. FLOORSMAX_MODE：客戶所居建築物正規化後最高樓層眾數。
67. FLOORSMIN_MODE：客戶所居建築物正規化後最低樓層眾數。
68. LANDAREA_MODE：客戶所居建築物正規化後土地面積眾數。
69. LIVINGAPARTMENTS_MODE：客戶居住用建築物正規化後眾數。
70. LIVINGAREA_MODE：客戶所居建築物正規化後居住區域面積眾數。
71. NONLIVINGAPARTMENTS_MODE：客戶非居住用建築物正規化後眾數。
72. NONLIVINGAREA_MODE：客戶所居建築物正規化後非居住區域面積眾數。
73. APARTMENTS_MEDI：客戶所居建築物正規化後的面積中位數。
74. BASEMENTAREA_MEDI：客戶所居建築物地下室正規化後的面積中位數。
75. YEARS_BEGINEXPLUATATION_MEDI：客戶所居建築物正規化後施工年數中位數。
76. YEARS_BUILD_MEDI：客戶所居建築物正規化後的建築年齡中位數。
77. COMMONAREA_MEDI：客戶所居建築物正規化後公共區域面積中位數。
78. ELEVATORS_MEDI：客戶所居建築物正規化後電梯數量中位數。
79. ENTRANCES_MEDI：客戶所居建築物正規化正規化後出入口數量中位數。
80. FLOORSMAX_MEDI：客戶所居建築物正規化後最高樓層中位數。
81. FLOORSMIN_MEDI：客戶所居建築物正規化後最低樓層中位數。
82. LANDAREA_MEDI：客戶所居建築物正規化後土地面積中位數。

83. LIVINGAPARTMENTS_MEDI：客戶居住用建築物正規化後中位數。
84. LIVINGAREA_MEDI：客戶所居建築物正規化後居住區域面積中位數。
85. NONLIVINGAPARTMENTS_MEDI：客戶非居住用建築物正規化後中位數。
86. NONLIVINGAREA_MEDI：客戶非居住用建築物正規化後居住區域面積中位數。
87. FONDKAPREMONT_MODE：檢修基金帳戶類型。
88. HOUSETYPE_MODE：客戶所居建築物型態。
89. TOTALAREA_MODE：客戶所居建築物經正規化後面積眾數。
90. WALLSMATERIAL_MODE：客戶所居建築物牆壁材質之眾數。
91. EMERGENCYSTATE_MODE：客戶所居建築物是否有緊急出口。
92. *OBS_30_CNT_SOCIAL_CIRCLE：客戶的社交環境中有多少次觀察到 30 天過期的貸款情況。
93. *DEF_30_CNT_SOCIAL_CIRCLE：客戶的社交環境中有多少次觀察到 30 天內未按時還款的貸款情況。
94. *OBS_60_CNT_SOCIAL_CIRCLE：客戶的社交環境中有多少次觀察到 60 天過期的貸款情況。
95. *DEF_60_CNT_SOCIAL_CIRCLE：客戶的社交環境中有多少次觀察到 60 天內未按時還款的貸款情況。
96. *DAYS_LAST_PHONE_CHANGE：客戶在申請貸款前多少天換過手機。
97. *FLAG_DOCUMENT_2：是否有填文件 2 資料。
98. *FLAG_DOCUMENT_3：是否有填文件 3 資料。
99. *FLAG_DOCUMENT_4：是否有填文件 4 資料。
100. *FLAG_DOCUMENT_5：是否有填文件 5 資料。
101. *FLAG_DOCUMENT_6：是否有填文件 6 資料。
102. *FLAG_DOCUMENT_7：是否有填文件 7 資料。

103. *FLAG_DOCUMENT_8：是否有填文件 8 資料。

104. *FLAG_DOCUMENT_9：是否有填文件 9 資料。

105. *FLAG_DOCUMENT_10：是否有填文件 10 資料。

106. *FLAG_DOCUMENT_11：是否有填文件 11 資料。

107. *FLAG_DOCUMENT_12：是否有填文件 12 資料。

108. *FLAG_DOCUMENT_13：是否有填文件 13 資料。

109. *FLAG_DOCUMENT_14：是否有填文件 14 資料。

110. *FLAG_DOCUMENT_15：是否有填文件 15 資料。

111. *FLAG_DOCUMENT_16：是否有填文件 16 資料。

112. *FLAG_DOCUMENT_17：是否有填文件 17 資料。

113. *FLAG_DOCUMENT_18：是否有填文件 18 資料。

114. *FLAG_DOCUMENT_19：是否有填文件 19 資料。

115. *FLAG_DOCUMENT_20：是否有填文件 20 資料。

116. *FLAG_DOCUMENT_21：是否有填文件 21 資料。

117. *AMT_REQ_CREDIT_BUREAU_HOUR：客戶提交申請的前一小時銀行查詢客戶資料次數。

118. *AMT_REQ_CREDIT_BUREAU_DAY：客戶提交申請的前一天銀行查詢客戶資料次數（未包含前一小時）。

119. *AMT_REQ_CREDIT_BUREAU_WEEK：客戶提交申請的前一週銀行查詢客戶資料次數（未包含前一天）。

120. *AMT_REQ_CREDIT_BUREAU_MON：客戶提交申請的前一個月銀行查詢客戶資料次數（未包含前一週）。

121. *AMT_REQ_CREDIT_BUREAU_QRT：客戶提交申請的前一季銀行查詢客戶資料次數（未包含前一月）。

122. *AMT_REQ_CREDIT_BUREAU_YEAR：客戶提交申請的前一年銀行查詢客戶資料次數（未包含前一季）。

新增變數：

1. SUM_FLAG_DOCUMENT：客戶填了多少文件。
2. SUM_AMT_REQ_CREDIT_BUREAU：客戶提交申請前一整年查詢客戶資料次數。
3. missing_ratio：遺失值比例。

表 9: 遺失值大於 32% 的變數

變數名稱	遺失值筆數	百分比
COMMONAREA_AVG	214229	69.87
COMMONAREA_MODE	214229	69.87
COMMONAREA_MEDI	214229	69.87
NONLIVINGAPARTMENTS_AVG	212872	69.43
NONLIVINGAPARTMENTS_MODE	212872	69.43
NONLIVINGAPARTMENTS_MEDI	212872	69.43
FONDKAPREMONT_MODE	209671	68.38
LIVINGAPARTMENTS_AVG	209567	68.35
LIVINGAPARTMENTS_MODE	209567	68.35
LIVINGAPARTMENTS_MEDI	209567	68.35
FLOORSMIN_AVG	208024	67.85
FLOORSMIN_MODE	208024	67.85
FLOORSMIN_MEDI	208024	67.85
YEARS_BUILD_AVG	203883	66.50
YEARS_BUILD_MODE	203883	66.50
YEARS_BUILD_MEDI	203883	66.50
OWN_CAR_AGE	202330	65.99
LANDAREA_AVG	182027	65.37
LANDAREA_MODE	182027	65.37
LANDAREA_MEDI	182027	65.37
BASEMENTAREA_AVG	179395	58.51
BASEMENTAREA_MODE	179395	58.51

續接下頁

表 9 (續): 遺失值大於 32% 的變數

變數名稱	遺失值筆數	百分比
BASEMENTAREA_MEDI	179395	58.51
EXT_SOURCE_1	172886	56.39
NONLIVINGAREA_AVG	169161	55.17
NONLIVINGAREA_MODE	169161	55.17
NONLIVINGAREA_MEDI	169161	55.17
ELEVATORS_AVG	163381	53.29
ELEVATORS_MODE	163381	53.29
ELEVATORS_MEDI	163381	53.29
WALLSMATERIAL_MODE	155859	50.83
APARTMENTS_AVG	155583	50.74
APARTMENTS_MODE	155583	50.74
APARTMENTS_MEDI	155583	50.74
ENTRANCES_AVG	154352	50.34
ENTRANCES_MODE	154352	50.34
ENTRANCES_MEDI	154352	50.34
LIVINGAREA_AVG	153880	50.19
LIVINGAREA_MODE	153880	50.19
LIVINGAREA_MEDI	153880	50.19
HOUSETYPE_MODE	153821	50.17
FLOORSMAX_AVG	152550	49.75
FLOORSMAX_MODE	152550	49.75
FLOORSMAX_MEDI	152550	49.75
YEARS_BEGINEXPLUATATION_AVG	149550	48.78
YEARS_BEGINEXPLUATATION_MODE	149550	48.78
YEARS_BEGINEXPLUATATION_MEDI	149550	48.78
TOTALAREA_MODE	147978	48.26
EMERGENCYSTATE_MODE	145315	47.39

表 10: 交叉驗證前被去除變數與目標變數的 Mutual Information

變數名稱	Raw Data	UnderSampling	OverSampling
CODE-GENDER	0.001455	0.004762	0.004725
NAME-INCOME-TYPE	0.002115	0.007182	0.007373
FLAG-MOBIL	0	0	0.000001
NAME-FAMILY-STATUS	0.000809	0.002415	0.002642
FLAG-DOCUMENT-2	0.000009	0.00002	0.000032
FLAG-DOCUMENT-3	0.001005	0.003616	0.003497
FLAG-DOCUMENT-4	0.000007	0.000028	0.000031
FLAG-DOCUMENT-5	0	0.000005	0
FLAG-DOCUMENT-6	0.000458	0.001731	0.001762
FLAG-DOCUMENT-7	0.000001	0.000024	0.000003
FLAG-DOCUMENT-8	0.000035	0.00008	0.00011
FLAG-DOCUMENT-9	0.00001	0.000037	0.00003
FLAG-DOCUMENT-10	0.000002	0.000028	0.000009
FLAG-DOCUMENT-11	0.00001	0.000029	0.000039
FLAG-DOCUMENT-12	0	0	0.000002
FLAG-DOCUMENT-13	0.000092	0.0003	0.000358
FLAG-DOCUMENT-14	0.000057	0.000128	0.000223
FLAG-DOCUMENT-15	0.000028	0.000133	0.000108
FLAG-DOCUMENT-16	0.000077	0.00024	0.000293
FLAG-DOCUMENT-17	0.000008	0.000013	0.000026
FLAG-DOCUMENT-18	0.000035	0.000128	0.000105
FLAG-DOCUMENT-19	0	0.000006	0.000002
FLAG-DOCUMENT-20	0	0	0
FLAG-DOCUMENT-21	0.000006	0.000033	0.000011

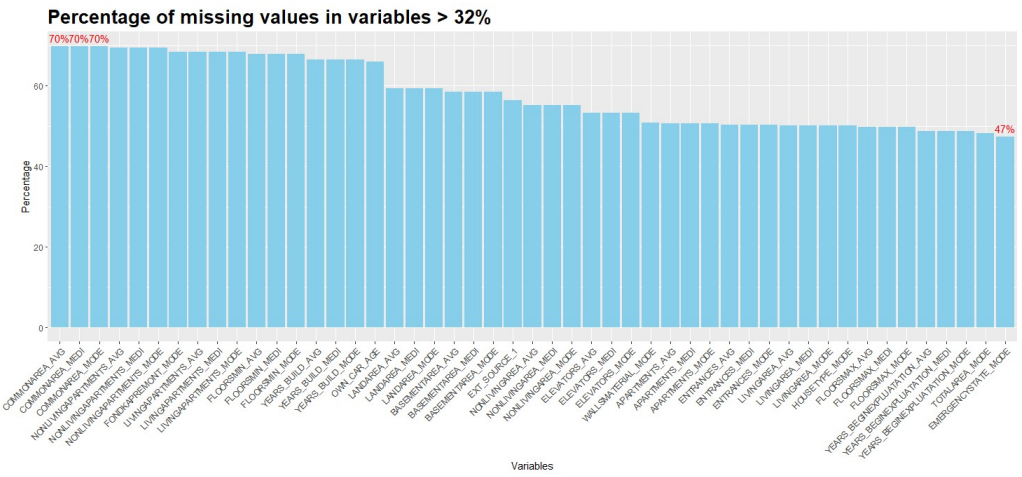


圖 31: 超過 32 % 遺失值的變數

表 11: 類別型變數 (TARGET) VS 數值型變數

數值型變數	p-value	Correlation	名次
EXT-SOURCE-2	0	-0.268909	1
EXT-SOURCE-3	0	-0.242487	2
YEARS-BIRTH	0	-0.137025	3
YEARS-LAST-PHONE-CHANGE	0	-0.10221	4
YEARS-ID-PUBLISH	0	-0.093831	5
YEARS-REGISTRATION	0	-0.079577	6
AMT-GOODS-PRICE	0	-0.07716	7
REGION-POPULATION-RELATIVE	0	-0.07172	8
AMT-CREDIT	0	-0.058562	9
DEF-30-CNT-SOCIAL-CIRCLE	0	0.055519	10
DEF-60-CNT-SOCIAL-CIRCLE	0	0.05343	11
HOUR-APPR-PROCESS-START	0	-0.044562	12
MISSING-RATIO	0	0.041493	13
CNT-CHILDREN	0	0.036886	14
AMT-REQ-CREDIT-BUREAU-HOUR	0	0.03473	15
SUM-FLAG-DOCUMENT	0	0.030818	16
AMT-ANNUITY	0	-0.025253	17
SUM-AMT-REQ-CREDIT-BUREAU	0	0.021073	18
CNT-FAM-MEMBERS	0	0.018427	19
OBS-30-CNT-SOCIAL-CIRCLE	0	0.01607	20
OBS-60-CNT-SOCIAL-CIRCLE	0	0.015858	21
AMT-REQ-CREDIT-BUREAU-WEEK	0	-0.014009	22
AMT-REQ-CREDIT-BUREAU-QRT	0.000001	0.00647	23
AMT-INCOME-TOTAL	0.00019	-0.004976	24
AMT-REQ-CREDIT-BUREAU-DAY	0.05339	-0.002577	25
AMT-REQ-CREDIT-BUREAU-MON	0.14999	-0.001921	26
AMT-REQ-CREDIT-BUREAU-YEAR	0.44256	0.001024	27

表 12: 工作分配

職位	工作內容
林貫原 (PM)	專案進度管理、決策管理、文件管理
許政揚 (副 PM)	協助進度管理、網頁雛形、小組報告
周昱宏 (DS)	資料清洗、資料分析
楊廷紳 (DS)	程式管理、資料分析
留筠雅 (SD)	系統架構、介面架構、使用說明書撰寫
易祐辰 (TS)	資料視覺化、資料搜集

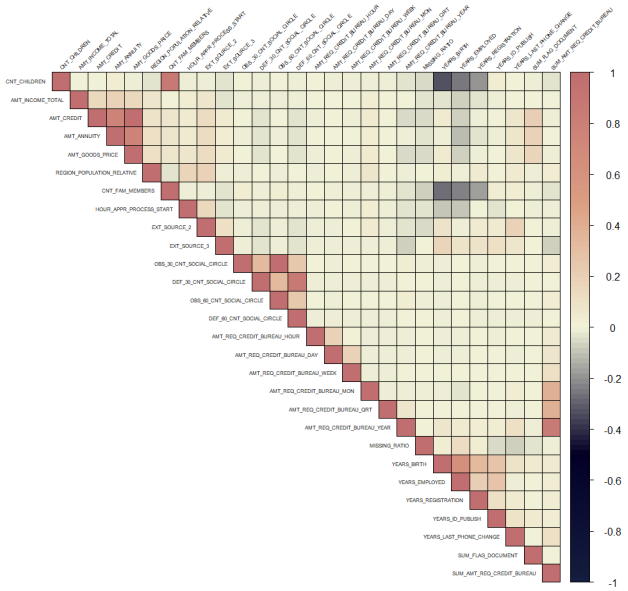


圖 32: 數值型變數相關係數圖

進度
112學年
3月5日
3月12日
3月19日
3月26日
4月2日
4月9日
4月16日
4月23日
4月30日
5月7日
5月14日
5月21日
5月28日
6月4日
討論職位分工
討論研究主題
資料前處理(遺失值插補)
探索資料分析
製作第一次小組簡報
1st Group Project Presentation
預測模型設計
模型效能驗證與調整
模型結果視覺化
製作第二次小組簡報
2nd Group Project Presentation
R Shiny 教學(撰寫網頁雛形)
成品展示
模擬期末報告、書面
期末報告

圖 33: 甘特圖