

臨門一腳隊_分析說明書

一、 資料處理與特徵選取

● Data Cleaning

異常值處理

將 NA 值過多之欄位刪除（參見附件一），忽略該變數不計，以免影響預測結果，造成誤判。其他未捨去之欄位 NA 值，我們利用 Mice 推估可能值並填補上去。

至於變數 CUST_9_SEGMENTS_CD 未選取之原因，是因為它是類別型數據，我們無法數字化它的類別項目，故不選取。

變數轉換

進行資料預處理時，我們先將 NA 數目過多者剔除，解決異常值問題。接下來，我們處理如：「低」、「中」、「中高」、「高」的類別類型資料，將類別型資料以無順序意義的數字來表示。

其次，我們先用 Excel 內建取代功能，把原始資料中的 Y（Yes）／N（No）分別轉換成 1／0，再利用 R 將屬於 Character 類型資料，轉換成 Number 或者 Integer，以免在後續跑訓練模型時可能會造成 Error。

我們將不同級距變數的數值改為不同公差的等差數列，如：「Age」、「APC_1ST_AGE」、「INSD_1ST_AGE」、「RFM_R」、「REBUY_TIMES_CNT」、「LIFE_CNT」。（轉換數值參考附件二）

數據缺失值填補

對變數 NA 值的處理，我們考慮「填補平均值」、「以 0 填補」，但可能會造成填補資訊無意義，或者造成資訊扭曲改

變，最後我們我們使用「MICE」套件建立 3 個訓練資料集，30 個 maxit，並使用 CART 決策樹，進行遺漏值預測，我們針對去除異常值後的 116 個變數建立填補好的資料集，以進行後續的建模。

● Data Processing

基礎資料分析

最開始我們利用敘述統計篩選出 17 個變數，作為線性回歸的數據。作法為：將符合預測目標的 ($Y1 == "Y"$) 篩選為新的 data，再從新 data 一個個檢測變數中也為 Yes 的。我們發現，利用敘述統計篩選出來的 17 個變數 Run 線性回歸的 Submit 初始分數到達約 0.792。因此，我們決定以線性回歸為主要改良方向，其餘時間也不斷測試有沒有表現更好的模型。但後來利用 R 提供的逐步回歸來計算 step() 函數，選出最小的 AIC 信息統計量逐步刪減項目，我們發現剔除 OCCUPATION_CLASS_CD 後的線性回歸更為貼合。

產生新變數過程說明

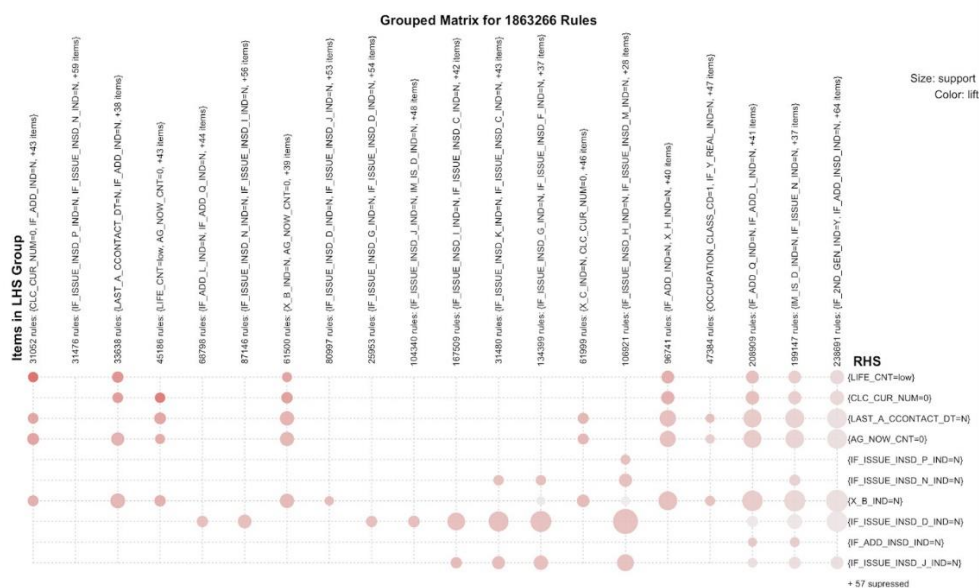
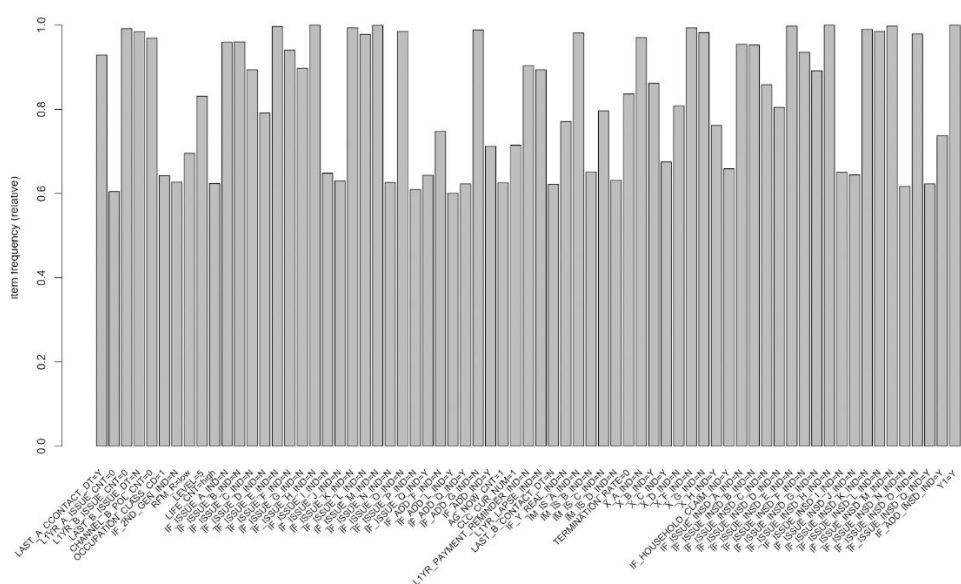
敘述統計：

以敘述統計的方式選出 17 個對 Y1 影響較關聯的變數。方法是以 R 程式開啟，先從原始資料中篩出符合我們預測期待結果 ($Y1 == "Y"$) 的資料，再從篩出的資料中一一篩各個變數與 $Y1 == "Y"$ 的關係，

關聯規則：

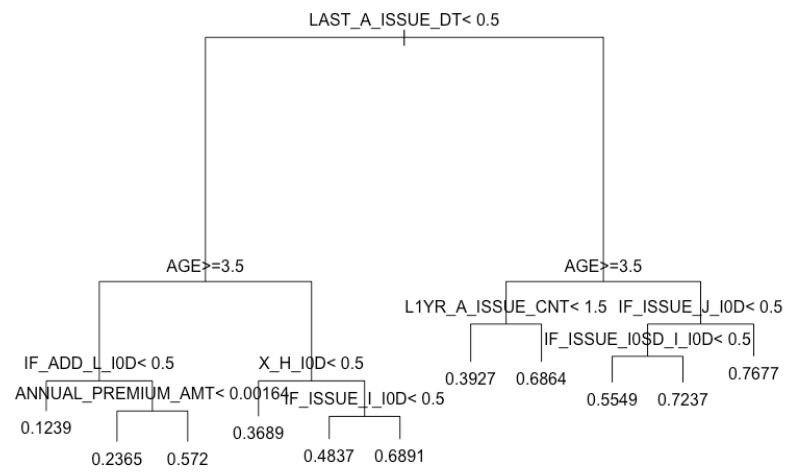
使用 Apriori 演算法是找出高頻項之布林值關聯規則中最具代表性的演算法。他採用水平方向進行項目集的搜尋。若某選項集為高頻，則其所有子集合必定是高頻項目集。

一開始先以 0.1 最小支持度與 0.4 最小信賴門檻逐步增加到 0.5 的最小支持度與 0.9 的信賴門檻來產生的關聯規則，最後將塞選出來的變數聯集敘述統計的變數，最後產生 43 個變數。



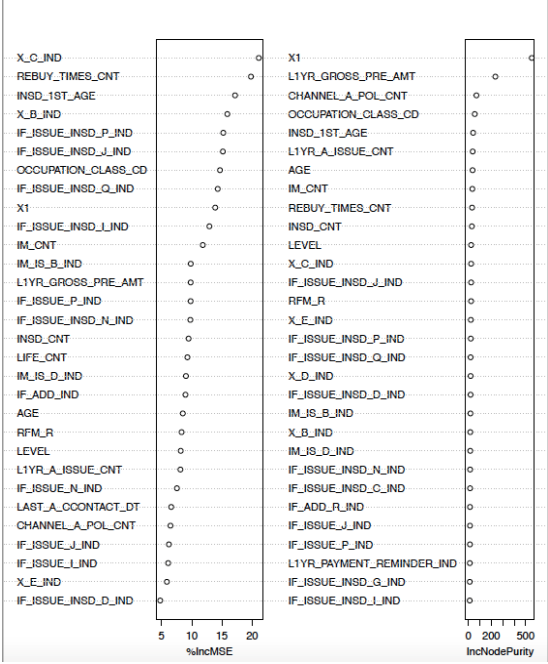
決策樹：

利用 **Cart** 套件訓練模型。並且把 **cp** 直設在 0.003 以免過度配適。



隨機森林：

比起決策樹，隨機森林更能有效地防止過度配適，所以我們利用隨機森林評估變數重要性，並使用套件（"varImpPlot"）函數畫圖，以「正確率下降指標」以及「分支不純度降低指標」來排序變數重要性，以下篩選排列出 38 個變數。



新增二次項與三次項：

因為二元項目（1 或 0）的平方項仍為 1 或 0，因此我們在交集項中選取不為二元（1 或 0）的變數數據資料做乘二次項與三次項來觀察。

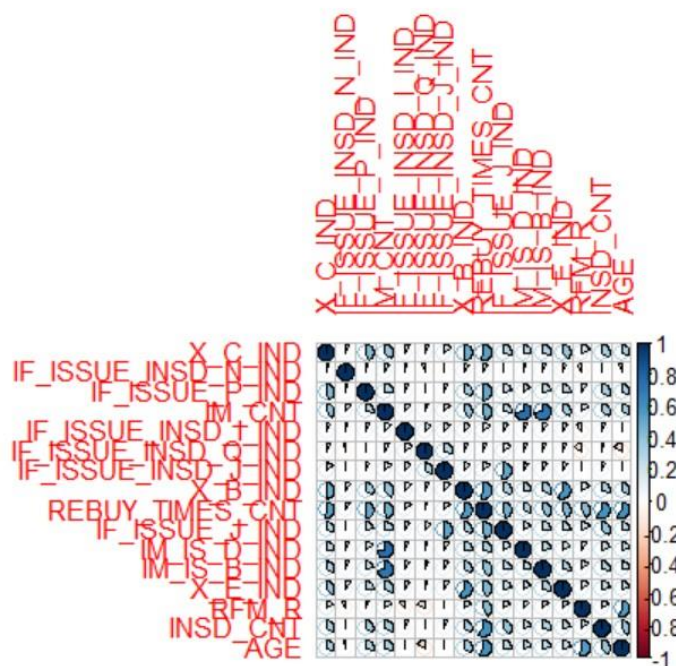
找交集：

我們使用 Random Forest 跑出來的重要變數取交集來檢視「正確率下降指標」以及「分支不純度降低指標」兩種指標的變數解釋程度。

交乘項：

我們利用 Corrplot 的方法（method：pie）使用 random forest 的交集來繪製 Corrplot，再將大於半圓的 Pie 找出兩變數相關性顯著者來新增為交乘項。

```
corrplot::corrplot(m, method = "Pie")
```



● Feature Selection

重要變數選擇過程

最初是用敘述統計，加上 Random forest 交集後項目，再利用 step 逐步刪減項目，使線性回歸準確率提高並貼合。

重要變數與預測變數之關聯說明

在嘗試 Log 指數模型與 LASSO 模型後，我們發現線性回歸仍最為貼合預測目標，因此重要變數與預測變數應為線性線性關係，且由於加入三次項的 AGE 之後，預測結果特為貼合、顯著，因此我們認為 AGE 也是影響 Y1 的重要變數。

二、模型選擇與驗證成效說明

● Model Selection

模型選擇與驗證過程

我們一開始用敘述統計選出的 17 個變數來跑單一變數的線性回歸（Linear regression），之後嘗試 Log 與 LASSO，但變數不確定較大，擔心效果不彰，因此先繼續優化線性回歸，於 17 個變數中加入使用關聯規則後挑出的變數，增加為 43 個變數。因為是用線性回歸於是我們嘗試增加變數項來讓線更符合預測值，然而在增減的過程中 43 個變數已經達到了極限。往後我們又增加了隨機森林選擇的變數，一系列的增增減減後再加上二次項、三次項、交乘項來跑複回歸。


除了線性之外，我們想瞭解自變數與應變數之間的關係是否可能為非線性關係，所以使用非線性的 Log 指數模型。但是發現 Log 指數的結果反而較線性回歸差，因此我們決定捨棄 Log 指數模型不用。

再來嘗試的是適用於二元分類的 LASSO 模型，由於本數據的二元資料，理論上非常適合應用，但 LASSO 需要 Test 與 Train 都是相同矩陣大小，但我們的數據 Test 為 150,000 筆、Train 則是 100,000 筆資料，矩陣大小不一，為了解決這個問題，我們先將 Test 分割成 100,000 和 50,000 筆資料與 Train 跑，而為了配合切割後的 50,000 筆 Test 資料，又再將 Train 隨機挑出 50,000 筆資料跑 LASSO 模型，因為是隨機挑出，我們無法保證是否選入過多偏差值，所以效果十分糟糕，因此我們最終仍使用線性回歸而非 LASSO 模型。

模型成效說明

線性回歸：


1. 58 個變數

 submit_test.csv	9/23/2019 11:36:00 PM	0.7902700508
2019/9/23 58 var linear reg		

2. 43 個變數（比 58 個變數優）

 submit_test.csv	9/27/2019 2:41:38 PM	0.8122606024
2019/9/27 43_2 var linear		

3. 43 個變數、二次、三次項與交乘項（最佳）

 submit_test.csv	9/30/2019 12:08:01 AM	0.8343019831
2019 / 9/ 30 tri 43 binary reg		

LASSO：

 submit_test_ridge_58.csv	10/4/2019 4:40:34 PM	0.4333381252
ridge regression 58		

Log：

 submit_test.csv	9/20/2019 11:42:28 AM	0.4761889137
2019/9/20 log binomial reg (re_train)		

綜上述所言，我們共嘗試了三個模型，但只有線性回歸的複回歸最為顯著，因此，在此之後，我們都集中探討如何優化線性回歸模型。最佳表現結果的是有利用 Mice 填補缺漏值，並且包含二次、三次項與交乘項的線性回歸複回歸模型。

附件：

附件一：（A_IND、B_IND、C_IND、L1YR_C_CNT、
DIEBENEFIT_AMT、DIEACCIDENT_AMT、
POLICY_VALUE_AMT、ANNUITY_AMT、EXPIRATION_AMT、
ACCIDENT_HOSPITAL_REC_AMT、
DISEASES_HOSPITAL_REC_AMT、
OUTPATIENT_SURGERY_AMT、
INPATIENT_SURGERY_AMT、PAY_LIMIT_MED_MISC_AMT、
FIRST_CANCER_AMT、ILL_ACCELERATION_AMT、
ILL_ADDITIONAL_AMT、LONG_TERM_CARE_AMT、
MONTHLY_CARE_AMT、IF_HOUSEHOLD_CLAIM_IND、
LIFE_INSD_CNT、IF_ISSUE_INSD_A_IND、
IF_ISSUE_INSD_B_IND、IF_ADD_INSD_F_IND、
IF_ADD_INSD_L_IND、IF_ADD_INSD_Q_IND、
IF_ADD_INSD_G_IND、IF_ADD_INSD_R_IND、
FINANCETOOLS_A、FINANCETOOLS_B、
FINANCETOOLS_C、FINANCETOOLS_D、
FINANCETOOLS_E、FINANCETOOLS_F、
FINANCETOOLS_G）

附件二：

AGE

low = 20

mid = 35

midhigh = 50

high = 65

APC_1ST_AGE

low = 22

mid = 33

midhigh = 44

high = 55

INSD_1ST_AGE

low = 0

mid = 20

midhigh = 40

high = 60

RFM_R

low = 1

mid = 4

midhigh = 7

high = 10

REBUY_TIMES_CNT

low = 5

mid = 10

midhigh = 15

high = 20

LIFE_CNT

low = 5

mid = 15

midhigh = 25

high = 35