GOGOGO CWW 噴_分析說明書

一、資料處理與特徵選取

(一) Data Cleaning and Data Processing

Step 1：在一開始讀進資料時，有許多變數的變數類型不對，因此我們先調整各個變數的類型，Ex ： 數值型轉類別型。

Step 2 ： 觀察 NA 值、填補 NA 值

1. 具有相同 NA 值的變數

首先，我們把資料集“train.csv”及“test.csv”合併，觀察整筆資料 NA 值的分布狀況並對其做出猜想：

A_IND	B_IND	C_IND

我們觀察這三個變數具有 NA 的資料皆相同(各有 198492 筆)，因此我們猜想這些保戶是完全沒有接觸 A、B、C 三電子報，原因可能是沒有相關資訊來源，因為若是有了解相關資訊但不使用電子報，則應填寫選項「NO」，但其為缺失值，所以我們認為原因應為保戶未取得相關資訊，因此我們將此缺失值設定為一新變數值「None」。

FINANCETOOLS_A	FINANCETOOLS_B	FINANCETOOLS_C
FINANCETOOLS_D	FINANCETOOLS_E	FINANCETOOLS_F
FINANCETOOLS_G		

接著觀察到上述表格內變數具有 NA 的資料皆相同(各有 156162 筆)，因此我們猜想這些保戶是完全沒有接觸 A-G 等七個理財工具，原因亦為沒有相關資訊來源，因此我們也將此處的缺失值設定為「None」。

IF_ADD_INSD_F_IND	IF_ADD_INSD_L_IND
IF_ADD_INSD_G_IND	IF_ADD_INSD_R_IND
IF_ADD_INSD_Q_IND	

而此表格內為具有相同 NA 值的五個變數(各有 129347 筆)，對此我們猜測這些保戶是完全沒有接觸 F-R 等五附約，可能是曾填寫的保單無附約，或是未曾投保主約，因此也不會有連帶的附約資訊，導致資料呈現缺失值，而我們將其轉為一新變數值「None」。

APC_1ST_AGE	APC_1ST_YEARDIF	REBUY_TIMES_CNT
RFM_M_LEVEL	TERMINATION_RATE	

同樣地，這五個變數具有 NA 的資料也相同(各有 107496 筆)，因此我們猜想這些保戶在「未擔任過要保人」的情況下，「第一次當要保人距今的時間」也會
缺失，也不會有「再次購買保險」以及「解約」的行為，因為未曾購買保險導致「曾投保主約件數」也不會存在，所以我們將這些缺失值令為「None」。

DIEBENEFIT_AMT	DIEACCIDENT_AMT
ANNUITY_AMT	DISEASES_HOSPITAL_REC_AMT
EXPIRATION_AMT	ACCIDENT_HOSPITAL_REC_AMT
POLICY_VALUE_AMT	OUTPATIENT_SURGERY_AMT
LONG_TERM_CARE_AMT	ILL_ACCELERATION_AMT
FIRST_CANCER_AMT	PAY_LIMIT_MED_MISC_AMT
ILL_ADDITIONAL_AMT	INPATIENT_SURGERY_AMT
MONTHLY_CARE_AMT	

而這 15 個經過神秘轉換的變數擁有的缺失值個數皆相同（各有 69005 筆），雖然數字可能代表的真正意義我們並不確定，但我們認為這些數字可能為契約訂立時所保的金額，而保戶可能沒有在當年度保障項目投保這 15 個項目，才會產生缺失值，由於這些變數為數值型變數，與前面類別型變數不同，因此我們將其直接轉換為「0」，而不是「None」。

IF_ISSUE_INSD_A_IND	IF_ISSUE_INSD_B_IND
IF_ISSUE_INSD_C_IND	IF_ISSUE_INSD_D_IND
IF_ISSUE_INSD_E_IND	IF_ISSUE_INSD_F_IND
IF_ISSUE_INSD_G_IND	IF_ISSUE_INSD_H_IND
IF_ISSUE_INSD_I_IND	IF_ISSUE_INSD_J_IND
IF_ISSUE_INSD_K_IND	IF_ISSUE_INSD_L_IND
IF_ISSUE_INSD_M_IND	IF_ISSUE_INSD_N_IND
IF_ISSUE_INSD_O_IND	IF_ISSUE_INSD_P_IND
IF_ISSUE_INSD_Q_IND	

在此處，這 17 項主約變數（各有 49994 筆）與前面的 5 項附約變數不同的是：我們懷疑這些保戶可能對於壽險較無需求或因經濟等

其他因素而不進行投保，而非主約本身不存在或是受其他契約的影響，因此最後使資料表中呈現缺失值的狀態，而我們將其同樣表示為「None」。

INSD_1ST_AGE	INSD_LAST_YEARDIF_CNT
IF_ADD_INSD_IND	

在接下來的這個部分（此處三變數亦各有 185 筆相同的 NA 資料），如同要保人的分析方式，因為沒有「擔任過被保人」，因此也不會有後續的「被保相關資訊」，所以我們也將缺失值設定為「None」。

X_A_IND	X_B_IND	X_C_IND
X_D_IND	X_E_IND	X_F_IND
X_G_IND	X_H_IND	

最後，我們認為在這 8 個項目全為缺失值的保戶（各有 52 筆）可能如同「電子報服務」的部分，保戶亦完全不知道這些相關服務的存在，因此無法申辦各類型服務，才導致缺失值的產生，而非填選選項「No」，對此我們如同電子報，將缺失值表示為「None」。

2. 沒有同樣 NA 值的變數

(1) L1YR_C_CNT

首先，我們觀察到「近一年到 C 通路申辦次數」這個變數的缺失值共有 22 萬多筆，因此懷疑是這項變數在被填入時可能是保戶無使用，或是保戶可能未接觸，但兩者意義相近，所以我們將其缺失值直接設為「0」。

(2) ANNUAL_PREMIUM_AMT

在「年繳化保費」的部分，其缺失值的意涵可能為保戶可能無繳交保費，因此在資料輸入時因為保戶未填寫此欄位，造成該筆資料變為缺失值，所以我們認為其缺失值與無繳保費（欄位值為 0）意義相同，所以我們將缺失值改為「0」。

(3) RFM_R

在「上次要保人身分投保距今間隔時間」這項項目中，我們觀察到此變數與上面第一點中的「要保人」相關的變數，兩者的缺失值不同的僅有 25 筆，因此我們推斷如同第一點中提及的，這些人可能未曾成為過要保人，因此在「RFM_R」這個變數中就會變成缺失值，因此我們將其填為「None」。

(4) LEVEL

而「往來關係等級」此項變數擁有的缺失值各數與

「RFM_R」變數相同，因此我們也懷疑因為未曾擔任過要保人，自然不會與保險公司有這方面的接觸，因此我們也將其缺失值更改為「None」。

(5) ANNUAL_INCOME_AMT

「年收入」這項變數的缺失值無法填為「None」，因為每位客戶都必定擁有各自的年收入，且此變數為連續型變數，因此我們認為使用該欄位的「中位數」來填補缺失值最為適當。

(6) BMI

「保戶的 BMI 值」如同「年收入」是一項不可能值為 0 的變數，因此在這個變數的缺失值部分，我們亦使用「中位數」來填補缺失值（不使用「平均數」的原因是因為「平均數」易受到離群值的影響）。

(7) EDUCATION_CD

在「教育程度」這項變數中，雖然無法觀察到規律，且也非連續型變數因此無法使用「中位數」進行缺失值填補，但我們認為其缺失值是可以使用「眾數」去進行填補的，因此我們將其缺失值都以眾數「1」進行填補。

(8) OCCUPATION_CLASS_CD

對於「客戶職業類別對核保風險程度」這項變數，如同「教

育程度」，因此我們認為其缺失值是可以使用「眾數」去進行填補的，而將其缺失值都以眾數「3」進行填補。

(9) MARRIAGE_CD、GENDER

在「婚姻狀況」及「性別」這兩個變數，我們並沒有觀察出其中的規律，且也不適合使用「中位數」、「眾數」等進行缺失值的填補，因此最後我們使用 R 裡面的「mice 套件」來對這兩個變數的缺失值進行處理，而 mice 裡面方法的選擇則選用「CART 決策樹」來對缺失值進行預測。

Step 3 ： 改變資料型態

針對上述非數值型態的變數，我們將其中變數值有包含「None」的變數，我們認為「None」的含意與「NO」相近，因此我們也把「None」改變成「0」的形式；在剩下的非數值型變數（例如：年齡、地區、客群等），以「低→高」、「A1→E」、「A→H」之排序依序從 0 開始編號。至此我們已經把所有非數值型態的變數轉變成以數字表示的類別變數。

（二）Feature Selection

我們先對還沒經過神秘轉換的數值變數進行歸一化，然後採取 PCA 去對數值型變數降維，刪掉的依據是挑選了 15 個主成分，因

為可解釋變異已經達到 0.98，然後我再從這 15 個主成分裡面，看各個變數的權重來決定排序(總共排 1~20)。最後在看說哪幾個變數都沒有擠進過任何一個主成分的前 20 名，另外把出現過一次以及兩次的變數也去掉，因為出現頻率低而且出現時的權重也很小，所以也不考慮。我們選擇的變數有:

L1YR_A_ISSUE_CNT	BANK_NUMBER_CNT
L1YR_B_ISSUE_CNT	INSD_LAST_YEARDIF_CNT
CHANNEL_A_POL_CNT	BMI
CHANNEL_B_POL_CNT	OCCUPATION_CLASS_CD
IM_CNT	TERMINATION_RATE
APC_CNT	TOOL_VISIT_1YEAR_CNT
INSD_CNT	ACCIDENT_HOSPITAL_REC_AMT
APC_1ST_YEARDIF	DISEASES_HOSPITAL_REC_AMT
AG_CNT	OUTPATIENT_SURGERY_AMT
AG_NOW_CNT	INPATIENT_SURGERY_AMT
CLC_CUR_NUM	PAY_LIMIT_MED_MISC_AMT
L1YR_C_CNT	LIFE_INSD_CNT

還有我們也有比較從 xgboost 的 important 圖中也顯示了重要變數的排名，下圖是 xgboost 算出的前十重要變數:

	Feature	Gain	Cover	Frequency
2	INSD_LAST_YEARDIF_CNT	0.0816201099	0.0667428583	0.0694698355
7	DIEBENEFIT_AMT	0.0368433098	0.0452002555	0.0639853748
8	APC_1ST_YEARDIF	0.0326809754	0.0554647434	0.0589579525
9	DIEACCIDENT_AMT	0.0312923784	0.0368706865	0.0580438757
4	TOOL_VISIT_1YEAR_CNT	0.0534431893	0.0447464352	0.0374771481
1	CHANNEL_A_POL_CNT	0.1000676823	0.0369897359	0.0287934186
15	L1YR_GROSS_PRE_AMT	0.0180113745	0.0234235194	0.0287934186
14	BMI	0.0230614930	0.0316519439	0.0278793419
20	ANNUAL_PREMIUM_AMT	0.0119711837	0.0214235441	0.0278793419
21	LIFE_INSD_CNT	0.0117558542	0.0191778191	0.0265082267

那其實跟我們一開始判斷重要的變數幾乎是一樣的。

(三) 處理 Unbalanced 問題

因為 train 資料集在是否有購買重疾險商品(預測目標)上是不平衡的，No 跟 Yes 的比是 98000:2000。我們如果不對資料進行處理的話，那做出來的模型會傾向猜 No，因此我們採取了多種辦法去處理不平衡資料。我們採用了有:（1）smote 進行 oversampling 至 No 跟 Yes 的比是 98000:98000（2）smote 進行 undersampling 至 No 跟 Yes 的比是 2000:2000（3）oversampling 至 No 跟 Yes 的比是 98000:98000（4）undersampling 至 No 跟 Yes 的比是 2000:2000。(1)、(3)兩種 oversampling 方法效果都並不是很好，

因此我們主要採取(2)、(4)兩種 undersampling 的方法，但 undersampling 會讓我們有可能因為減少 No 的樣本而可能樣本會信息缺失，所以我採取了 EasyEnsemble 的方法去解決傳統隨機欠抽樣方法導致的信息缺失的問題。我們會做多次 undersampling 然後去比較模型結果來判斷 undersampling 後的樣本是否具有代表性。

二、模型選擇與驗證成效說明

Model Selection

(一) 模型選擇

我們一開始有使用隨機森林模型去分類，之後也有用 xgboost 模型去分類，xgboost 效果比隨機森林優，因此我們選擇了 xgboost。

(二) 訓練參數

我們以我們過去的經驗決定出參數的範圍，然後進行 5 次交叉驗證挑選最佳參數。各參數的範圍選擇如下：

參數	範圍選擇
max_depth	np.random.randint(6 , 30)
eta	np.random.uniform(0.01 , 0.3)
gamma	np.random.uniform(0.0 , 0.2)

subsample	np.random.uniform(0.6 , 1)
colsample_bytree	np.random.uniform(0.5 , 0.8)
min_child_weight	np.random.randint(1 , 41)
max_delta_step	np.random.randint(1 , 11)

最後我們選擇的參數是:

max_depth = 18

eta = 0.04567486016118333

nthread = 2

nrounds = 200

colsample_bytree = 0.5703148857513719

gamma = 0.16365455495795223

subsample = 0.7989452467573158

min_child_weight = 38

max_delta_step = 10

eval_metric = "auc"

eval_metric = "logloss"

objective = "binary:logistic"

(三) 訓練模型

我們原本的訓練集分成訓練集、驗證集，比例為 0.8 比 0.2，

訓練集選擇迭代 100 次到 1000 次之間，訓練出來的結果在驗證集上的表現都不錯而最後拿去測試的結果也都很穩定。

（四） 模型評估

我們以 train 每次迭代的 AUC 跟 logloss 去判斷模型的好壞，同時也必須避免 overfitting 的發生。每次不同的參數和 undersampling 資料的 AUC 都有達到 0.9 多，logloss 也都在 0.45 以下，因此我們認為模型是可以接受的。