IF 隊伍：2019 國泰大數據競賽分析說明書

一、 專案介紹

在保險業中，客戶購買保單之思考評估行為相當複雜且難以捉摸，對於銷售方面來說，業務員精準的行銷和維持良好的客戶關係，以及瞭解客戶本身的保險需求都相當重要，因此在本次專案中，我們想透過客戶過往的數據資料，預測在某個時間點中，公司既有客戶在未來三個月內是否會購買重疾險保單，以此建立一預測購買模型，希望能因此挖掘出保險需求較高的客戶，並提供這樣的方法給業務員去使用，讓業務員能瞄準這些較有潛在需求的客戶去做銷售，以期整體成本降低，且利益最大化。

二、 探索式分析 (EDA, Exploratory Data Analysis)

2.1 資料集描述

表 1

	資料筆數	欄位數
Train 資料集	100000筆	131
Test 資料集	150000筆	130

官方所提供之資料集中包含 Train 和 Test 兩種，如表 1 所示，在兩種資料集中，均將欄位「CUS_ID」作為索引值 (Index)，因此不納入欄位計算中，而在 Test 資料集中，缺乏欄位「Y1」，因為其為本專案欲求解的目標欄位。

2.2 Y1 欄位觀察

觀察 Y1 欄位之 N/Y 值次數累積圖，發現顧客是否有投保重疾險的分配極度不平均，Y 值(有投保)總計 2000 筆資料，N 值(沒投保)總計 98,000 筆資料，如圖 1 所示。

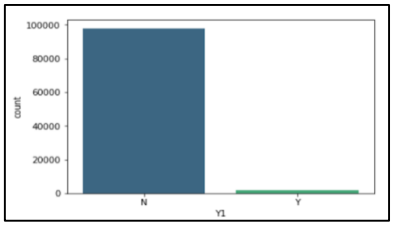

圖 1: Y1 欄位 N/Y 數量分布

2.3 遺漏值觀察

使用遺漏值解析圖觀察 Train 資料集與 Test 資料集中各個欄位的遺漏值分布狀況，如圖 2 所示，X 軸為欄位，Y 軸為顧客編號，有值的資料格以藍色表示，無值的資料格則以紅色表示。

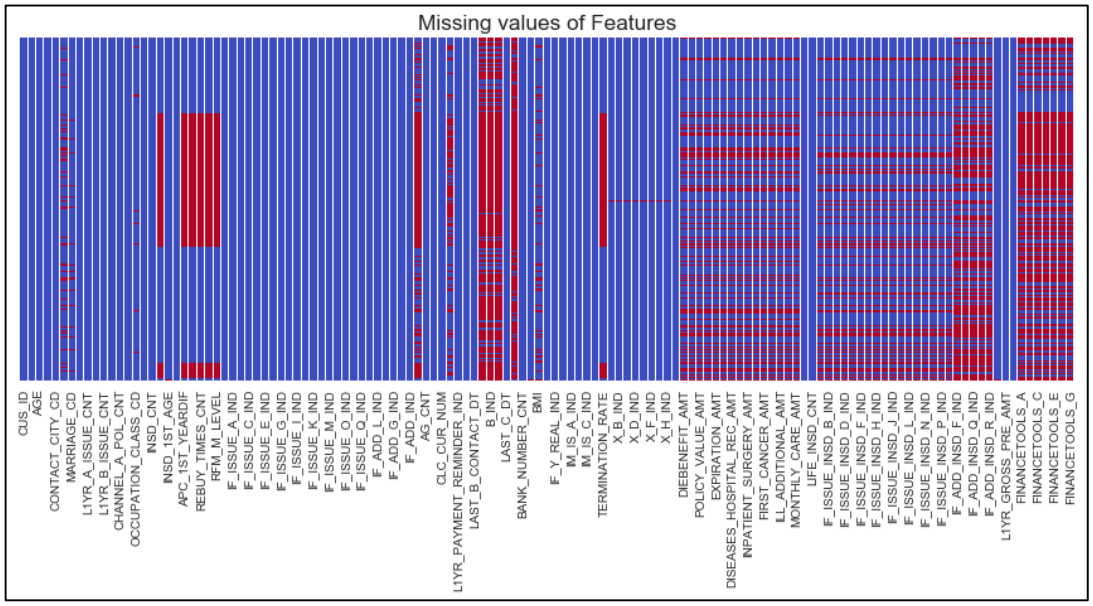

圖 2: 遺漏值解析圖

由此解析圖可觀察到 Train 資料集的遺漏值狀況為：(1)共有 73 個欄位有缺失值、(2)共 5783 個樣本其具缺失值的欄位將近所有欄位的一半、(3)在 131 個特徵中，共計 17 個欄位其遺漏值佔該欄位一半以上。Test 資料集的遺漏值狀況為共有 73 個欄位有缺失值，且 Test 資料集所具有缺失值的欄位與 Train 資料集所具缺失值的欄位完全相同。另外，可以發現解析圖上遺漏值的資料在部分欄位呈現平行的分布，同一顧客會同時在特定幾個欄位都有遺漏值，或是同時在特定幾個欄位都沒有遺漏值。

三、 資料清理與處理（Data Cleaning & Data Processing）

3.1 變數轉換

整合整份資料欄位值域的觀察，以及適配於選用的機器學習模型特性，規劃將類別型欄位資料轉換成數值型欄位資料，轉換規則如下：

(1) 低、中、中高、高 → 1 2 3 4

(2) 低、中、高 → 1 2 3

(3) N , Y → 0 1（含 label）

(4) F , M → 0 1

(5) "A1": 1, "A2": 2, "B1": 3, "B2": 4, "C1": 5, "C2": 6,"D": 7, "E": 8

(6) "A": 1, "B": 2, "C": 3, "D": 4, "E": 5,"F": 6, "G": 7, "H": 8

在此階段，將欄位「BMI 值」做離散化處理，分成是否在[0, 0.2)之間，是:1、否:0。透過核密度圖(參考圖 3)，可知在不同 BMI 值下會購買重疾險與不會購買重疾險的樣本數目分配。從圖中可觀察到當 BMI 值在 0~0.2 這個區間時，不買重疾險的樣本數明顯多於會買重疾險的樣本數；當 BMI 值不在 0~0.2 這個區間時，不買重疾險的樣本數明顯就少於會買重疾險的樣本數，或者兩者數量趨同。因此可以推論當某客戶 BMI 值落在 0~0.2 這個區間時，很有可能不會購買重疾險。由此得知 BMI 是否在[0,0.2)區間是解釋 Y1(是否購買重疾險)的重要特徵，因此將 BMI 值做離散化處理。

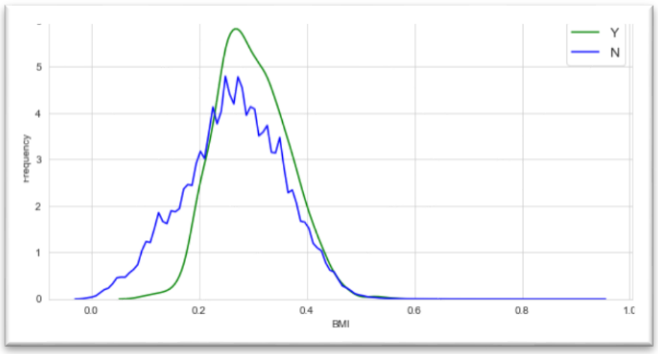

圖 3: BMI 分別在欄位 Y1 為 N 或 Y 之核密度圖

最後新增一個欄位用來適當描述顧客總年度的保障金額特性，新欄位為『TOTAL_AMT:總當年度保障』，計算方式為：當年度保障金

額相關欄位(共 15 欄)加總形成新的欄位，並且把該 15 欄刪除。

3.2 缺失值填補

3.2.1 填補缺失值方法順位:

一、 依照質性分析判斷結果來填補。

二、 回歸法填補(兩種方案):

方案一:

將資料集中沒有缺失值的欄位作為解釋變數，將欲填補的欄位作為被解釋變數，利用 ExtraTreesRegressor 模型中 feature_importances_功能，計算出每個解釋變數對於被解釋變數欄位的重要性分數，並依序排序，找出前 10 重要的欄位（解釋變數），再將該 10 個欄位分成: 第一重要欄位、第一 + 第二重要欄位、第一 + 第二 + 第三重要欄位……共 10 組欄位組合，依序用這 10 組特徵組合作為解釋變數，欲填補的欄位做為被解釋變數，用 ExtraTreesRegressor 模型進行運算，每組特徵運算時都會有一個正確率指標作為對應。若欲填補的欄位是類別型態變數，正確率指標為 F1 score；若欲填補的欄位是數值型態變數，正確率指標為 MSE 和 RMSE。最後取正確率指標最好的特徵組合，用 ExtraTreesRegressor 模型，來預測與填補欲填補的欄位的缺失值。如果每個正確率 F1 score 指標未達到 0.7，或 MSE 和 RMSE 不理想，則捨棄該方法。

方案二:

此方法為強化方案一更為謹慎的方法，如果方案一顯示的正確率指標結果理想，以求嚴謹，會考慮再用方案二來填補缺值。首先將沒有缺失值的欄位作為解釋變數，將欲填補的欄位作為被解釋變數，利用 ExtraTreesClassifier 模型，設定 10 種不同的種子，運算模型 10 次，每次運算都使用 feature_importances_功能計算出每個解釋變數的重要性分數，每次都列出前 10 重要的欄位（解釋變數），統計所有沒缺失值的欄位（解釋變數）出現在這 10 次運算中前 10 重要欄位中的頻率。最後取出頻率前幾高的欄位（解釋變數），用 ExtraTreesClassifier 模型，來預測與填補欲填補的欄位的缺失值。

三、 眾數與平均數填補:

將欲填補的欄位的缺失值以該欄位的眾數或平均數填補。

3.2.2 執行結果:

根據探索式分析 EDA 中的 2-3 節,「APC_1ST_AGE 首次擔任要保人年齡」、「REBUY_TIMES_CNT 再購次數」、「RFM_M_LEVEL 曾投保主約件數」、「RFM_R 上次要保人身份投保距今間隔時間」、「LEVEL 往來關係等級」(參考表 2)這些欄位的缺失值都代表這些樣本從未擔任過要保人,然而此缺失值的意涵在這五個欄位都不屬於任何一個分類類別,因此將這些欄位的缺失值都變成一個新的類別: 0。

表 2

REBUY_TIMES_CNT (再購次數)	資料筆數
1	27778
2	12267
3	6255
4	10418
NaN	43282

APC_1ST_AGE (首次擔任要保人年齡)	資料筆數
NaN	3620
3	1388
1	791
4	774
2	541

LEVEL (往來關係等級)	資料筆數
NaN	43305
5	28415
1	13940
4	6490
3	4041
2	3809

RFM_M_LEVEL (曾投保主約件數)	資料筆數
3	20799
5	11832
7	7296
8	7753
9	4865
10	4173
NaN	43282

RFM_R (上次要保人身份投保距今間隔時間)	資料筆數
1	17506
2	12291
3	14616
4	12293
NaN	43294

由探索式分析 EDA 中的 2-3 節,「APC_1ST_YEARDIF 首次成為要保人距今間隔時間」欄位的缺失值代表從未擔任過要保人,該缺失值的意涵不屬於任何欄位中的數值,然而因「APC_1ST_YEARDIF」為數值型變數,無法存在缺失值,因此將該欄位離散化,轉成名目尺度(參考表 3),並將缺失值變成一個新的類別。

表 3

APC_1ST_YEARDIF（首次成為要保人距今間隔時間）	
原本型態	離散化型態
NaN	0
[0,25]	1
(25,50]	2
(50,75]	3
(75,100]	4

表 4

TERMINATION_RATE(曾解約保單張數佔曾投保保單張數佔率)	
count	56718
mean	12.09
std	27.66
min	0
25%	0
50%	0
75%	0
max	100

承接探索式分析 EDA 中的 2-3 節,「ATION_RATE 曾解約保單張數佔曾投保保單張數佔率」欄位的缺失值代表從未擔任過要保人，而在這狀態下遺失值不屬於欄位裡任何數值，由於此欄位為數值型變數，無法存在缺失值，因此將該欄位離散化，並將缺失值變成一個新的類別，觀察表 4 欄位敘述統計可知，數值大多接近 0，因此直接將欄位轉換成兩種類別: 是 0、不是 0。而缺失值則獨立成一類。

欄位「OCCUPATION_CLASS_CD 客戶職業類別(各類別)對核保風險程度」採用回歸法方案一來填補。首先取出 15 個重要的解釋變數做 15 種組合，並將該組合依序作為解釋變數對「OCCUPATION_CLASS_CD 客戶職業類別(各類別)對核保風險程度」做預測，F1 score 大約都落在 0.69~0.71 間，最終選擇其中一種組合，由下列欄位作為解釋變數來預測填補缺失值: CONTACT_CITY_CD 、 TOOL_VISIT_1YEAR_CNT 、 LIFE_INSD_CNT 、 CUST_9_SEGMENTS_CD 、 AGE 、 L1YR_GROSS_PRE_AMT 、 CHARGE_CITY_CD 、 CHANNEL_A_POL_CNT、AG_CNT、AG_NOW_CNT。

欄位「ANNUAL_INCOME_AMT 年收入」則採用回歸法中結合方案一與方案二的做法來填補缺失值。首先透過方案一，取出前 25 名重要的欄位，並做 25 種組合，每種組合所顯示的正確率指標 MSE 大多都介於 1.9~2.3 之間，較難看出優劣之分。進一步採用方案二，設 50 種不同種子，做 50 次模型運算，其中「L1YR_GROSS_PRE_AMT 近一年實繳保費」、「AG_CNT 曾經經手過的業務員人數」、「AGE 年齡」、「LIFE_INSD_CNT 目前主約被保有效件數」、「CUST_9_SEGMENTS_CD 九大客群」該五個欄位出現在前 10 名特徵的行列出現 50~49 次，因此將該五項欄位作為解釋變數，預測並填補缺失值。

欄位「BMI 值」，承接 3.1 節變數轉換，將該欄位離散化，分成是否在[0, 0.2)之間，是:1、否:0，並將缺失值歸類為「否」這類。

資料集中「當年度保障金額相關欄位」(共 15 欄)，採用回歸法方案二來填補缺失值。先設 10 種不同種子，做 10 次運算，其中「AGE 年齡」、「IF_Y_REAL_IND 是否投保 Y 險」、「TOOL_VISIT_1YEAR_CNT 近一年業務員管理工具拜訪次數」、「IF_2ND_GEN_IND 是否為保戶二代」、「IF_S_REAL_IND 是否投保 S 險」、「IF_ADD_Q_IND 目前是否壽險保單持有有效類別_Q」、「APC_CNT 對應的要保人數」、「LIFE_INSD_CNT 目前主約被保有效件數」、「CONTACT_CITY_CD 聯絡地址_縣市」，該九個欄位出現在前 10 名特徵的行列皆出現 10 次，因此將該九項欄位作為解釋變數，預測並填補缺失值。

「目前是否壽險保單被保有效類別_A~R（主約）」(共 17 欄)的缺值可由「LIFE_INSD_CNT 目前主約被保有效件數(件)」欄位得知部分線索。如果樣本在該欄位的值為 0 時，代表該樣本當年沒有投保主約，理應在「目前是否壽險保單被保有效類別_A~R（主約）」這 17 個欄位中都為 0 （否）。然而將這 17 個欄位值對應「LIFE_INSD_CNT 目前主約被保有效件數(件)」欄位值為 0 時可以發現幾乎所有的缺失值都落在「LIFE_INSD_CNT」為 0 的情況。因此將這 17 個欄位的缺失值以 0（否)來填補。

觀察圖 4 可發現，當「OCCUPATION_CLASS_CD 客戶職業類別(各類別)對核保風險程度」為 1 時，女性數目明顯多於男性，因此推論當樣本在「OCCUPATION_CLASS_CD」欄位值為 1 時，該樣本很可能是女性。將「Gender 性別」欄位的缺失值對應到「OCCUPATION_CLASS_CD」欄位，如果該缺失值在「OCCUPATION_CLASS_CD」欄位的值為 1，則「Gender 性別」欄位的缺失值填補為 1（女性)，否則填補為 0（男性)。

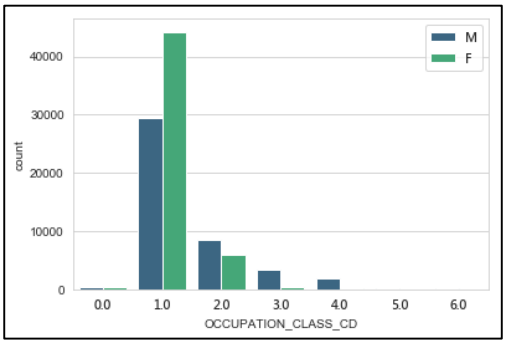

圖 4: OCCUPATION_CLASS_CD 欄位之男女(M/F)次數分配

從「L1YR_C_CNT 近一年到 C 通路申辦服務次數」欄位可知，如果三年沒有到 C 通路申辦服務代表近一年不會去 C 通路申辦服務，將「L1YR_C_CNT」的缺失值對應到「LAST_C_DT」欄位，如在「L1YR_C_CNT」欄位有缺失值的樣本在「LAST_C_DT」的欄位值為 0 時，該缺失值理應是 0，因此以 0 來填補。

資料集中「ANNUAL_PREMIUM_AMT」、「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」、「A_IND」、「B_IND」、「C_IND」、「FINANCETOOLS_A」、「FINANCETOOLS_B」、「FINANCETOOLS_C」、「FINANCETOOLS_D」、「FINANCETOOLS_E」、「FINANCETOOLS_F」、「FINANCETOOLS_G」這些欄位缺失值過多，直接予以刪除。

其他有遺失值的欄位中類別型態與數值型態的欄位分別以眾數、平均數填補。最後剩餘少數有缺失值的欄位，則直接將在該欄位有缺失的樣本刪除。

四、 模型建置

4.1 Further Processing

將多類別的名目變數做 One-Hot Encoding 處理，如果只有兩個類別的名目變數就已在資料清理時轉成 0、1 類別。多類別的名目變數為：CHARGE_CITY_CD、CONTACT_CITY_CD、MARRIAGE_CD、CUST_9_SEGMENTS_CD、APC_1ST_AGE、REBUY_TIMES_CNT、RFM_M_LEVEL、APC_1ST_YEARDIF、

TERMINATION_RATE、RFM_R、LEVEL。

4.2 Model Selection

因為比賽期間的時間限制與硬體設備限制，本團隊實較難為多個模型調整參數（Hyperparameter Tuning），以及在有限的上傳次數規則下，我們選擇採取最佳模型選用策略，僅針對成效較佳的兩個重要模型進行強化與上傳網站測試。

表 5

Model	AUC
Extreme Gradient Boosting (XGBoost)	0.8156
AdaBoostClassifier	0.8050
Random Forest	0.7979
Logistic Regression	0.7970
MLP	0.6635
Decision Tree	0.5264

分別將清理後的資料在 XGBoost、AdaBoost、Random Forest、DecisionTree、LogisticRegression、MLP 等六種模型運行，最終挑選兩個 AUC 最高的模型作為後續強化以及上傳測試的模型。上述六種模型都未調整參數，僅用預設的參數來運行。由表 5 可知，XGBoost 和 AdaBoost 的 AUC 位居前兩名。然而，由於前三名的模型都是以 Tree 為基底的模型，運算原理相近，為了後續上傳測試的模型能多樣化並增加分數進步的可能性，傾向選擇一個以 Tree 為和一個以非 Tree 為基底的模型；再者，位居第四的 Logistic Regression 模型 AUC 不至於落後太多，另外該模型還尚有許多優化的可能性，例如該模型本身容易受不平衡資料影響，後續可以透過重複抽樣(SMOTE Sampling)來改善，另外以 Tree 為基底的模型本身就有選擇重要特徵來運算的傾向，但 Logistic Regression 則不能，後續也可以透過特徵重要性選取(Feature Importance)來改善。因此，最終選擇 AUC 最高的 XGBoost 和非 Tree 為基底的 Logistic Regression 模型做為後續強化和上傳測試的模型。

4.3 Modeling

透過 Train_test_split 函式，將 Train 資料集分割成 67%的訓練資料(Training Data)和 33%的驗證資料(Validation Data)，後續都用那

67%的訓練資料來運行模型，33%來驗證模型表現。訓練資料的目標值為 Y1(是否購買重疾險)，預測值則為資料集中其餘欄位。模型運行完畢後則用該模型來預測 Test 資料集的目標值: Y1(是否購買重疾險)，並將該預測值上傳比賽平台，並得到一個初始分數。模型的部分採用由 4.2 節 Model Selection 所挑選的兩個模型: XGBoost、Logistic Regression。

針對 XGBoost 模型分別用五種不同的方式來運行，每次訓練完畢都會上傳比賽平台並得到一個初始分數: 1. 未做任何處理，模型參數直接採用其預設值、2. 調整參數、3. 調整參數並選擇重要特徵、4. 增加新的資料處理方式、5. 增加新的資料處理方式並調整參數；針對 Logistic Regression 模型分別用五種不同的方式來運行，每次訓練完畢都會上傳比賽平台並得到一個初始分數: 1. 未做任何處理，參數直接採用其預設值，直接上傳、2. 做 One-hot Encoding 並調整參數、3. 做 One-hot Encoding，調整參數並選擇重要特徵、4. 做 One-hot Encoding,調整參數並做重複抽樣(SMOTE Sampling)、5.增加新的資料處理方式。

針對 XGBoost 以第一種方式運行，未對其做任何處理，直接上傳。Train 資料集中 33%的驗證資料的驗證結果顯示 AUC 值為 0.815。模型的上傳成績的 AUC 值為 0.8469。

針對 XGBoost 以第二種方式運行，調整模型參數。調整的參數分別為：1. Learning Rate 運算值分別為 0.05, 0.1, 0.15, 0.2, 0.25、2. max_depth 運算值分別為 3, 5, 7, 9, 12、3. min_child_weight 運算值分別為 1, 3、4. gamma 運算值分別為 0.0, 0.1, 0.2、5. colsample_bytree 運算值分別為 0.3, 0.5, 0.7, 1。而最佳參數組合是 colsample_bytree 為 1，gamma 為 0，learning_rate 為 0.1， max_depth 為 3，min_child_weight 為 1。採用最佳的參數組合來運行模型後，Train 資料集 33%的驗證資料驗證結果顯示,AUC 值為 0.817。模型的上傳成績的 AUC 值為 0.8407。

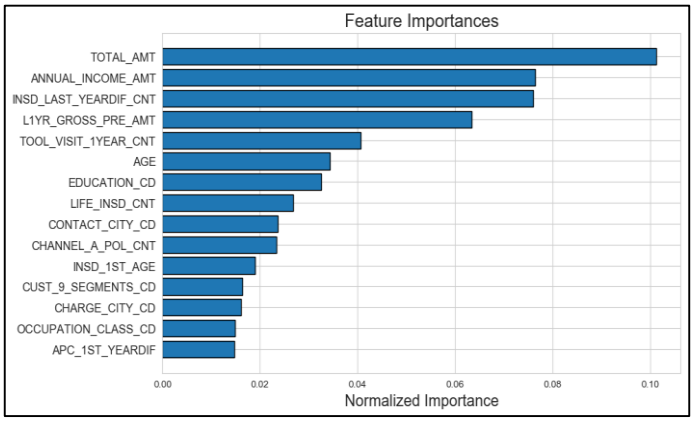

圖 5

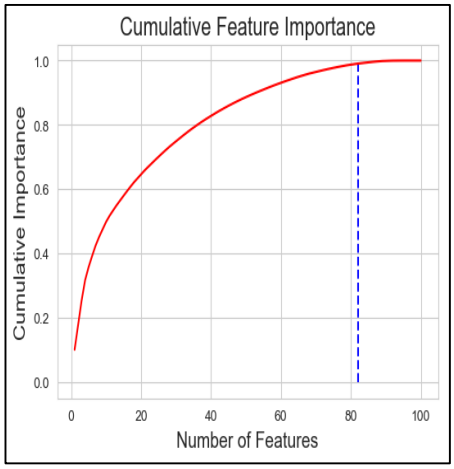

圖 6

針對 XGBoost 以第三種方式運行，選擇重要特徵並調整模型參數。重要特徵選擇的部分使用 FeatureSelector 套件，並將 feature importance 為 0、存在共線性的特徵給予刪除。其中 IF_ISSUE_INSD_H_IND 、IF_ISSUE_M_IND、IF_ISSUE_E_IND 在 feature importance 方面均為 0，均給予刪除；IF_ISSUE_INSD_H_IND 與 IF_ISSUE_H_IND 存在共線性，將其中一者刪除。由圖 5 則可得知所有特徵中對 Y1 欄位而言前 15 名重要的特徵，而圖 6 則顯示當特徵數達到 80 左右時幾乎已包含對 Y1 欄位而言所有的重要性。接著選用方法二的參數組合調整模型參數，並用最佳的參數來運行模型。Train 資料集 33%的驗證資料其驗證結果顯示，AUC 值為 0.818。模型的上傳成績的 AUC 值為 0.844。

針對 XGBoost 以第四種方式來運行，增加新的資料處理方式。新的資料處理方式並未把 AMT 系列 15 個欄位加總，每個欄位的缺失值都用中位數來填補；另外 AGE（年齡）與 IF_2ND_GEN_IND(是

否為保戶二代)兩個欄位合併為新欄位，新欄位的類別包含: 低年齡者且是保戶二代、中年齡者且不是保戶二代、中高年齡者且是保戶二代共三種類別。增加新的資料處理方式後，Train 資料集 33%的驗證資料其驗證結果的 AUC 值為 0.8143。模型的上傳成績的 AUC 值為 0.8440。

針對 XGBoost 以第五種方式來運行，增加新的資料處理方式並調整模型參數。新的資料處理方式和方法四相同，接著選用方法二的參數組合調整模型參數，並用最佳的參數來運行模型。增加新的資料處理方式並調整參數後，Train 資料集 33%的驗證資料其驗證結果的 AUC 值為 0.8217。模型的上傳成績的 AUC 值為 0.8427。

而針對 Logistic Regression 以第一種方式來運行，未對其做任何處理，直接上傳。Train 資料集 33%的驗證資料其驗證結果顯示，AUC 值為 0.796。模型的上傳成績的 AUC 值為 0.8321。

針對 Logistic Regression 以第二種方式來運行，做 One-hot Encoding 並調整模型參數。調整的參數分別為: 1. classifier__penalty 運算值分別為 l1, l2、2. classifier__C 運算值分別為 0.01, 0.1, 1, 10。而最佳參數組合是 classifier__penalty 為 l1，classifier__C 為 1。採用最佳的參數組合來運行模型後，Train 資料集 33%的驗證資料其驗證結果顯示，AUC 值為 0.801。模型的上傳成績的 AUC 值為 0.8329。

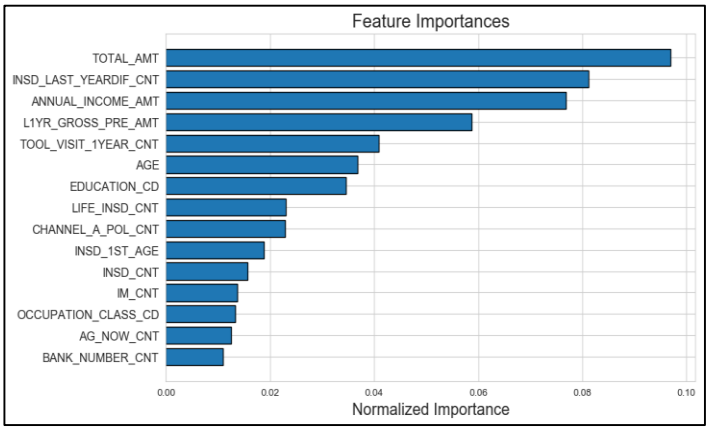

圖 7

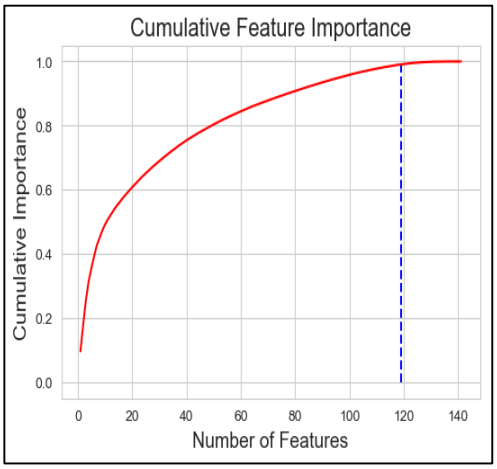

圖 8

針對 Logistic Regression 以第三種方式運行，做 One-hot Encoding、調整模型參數並選擇重要特徵。其中因為比 XGBoost 第三種方法多進行 One-hot Encoding 處理，因此在重要特徵選擇上也會有不同的結果。重要特徵選擇的部分使用 FeatureSelector 套件，將 feature importance 為 0、存在共線性的特徵給予刪除。其中 IF_ISSUE_INSD_H_IND 、 IF_ISSUE_M_IND 、 IF_ISSUE_INSD_E_IND、IF_ISSUE_E_IND 在 feature importance 方面均為 0，均給予刪除；IF_ISSUE_INSD_H_IND 與 IF_ISSUE_H_IND 存在共線性，將其中一者刪除。由圖 7 則可得知所有特徵中對 Y1 欄位而言前 15 名重要的特徵，而圖 8 則顯示當特徵數達到 120 左右幾乎已包含對 Y1 欄位而言所有的重要性。接著選用方法二的參數組合調整模型參數，並用最佳的參數來運行模型。做 One-hot Encoding、刪除不重要的特徵並採用最佳的參數組合，運行模型後，Train 資料集 33%的驗證資料其驗證結果顯示，AUC 值為 0.801。模型的上傳成績的 AUC 值為 0.8329。

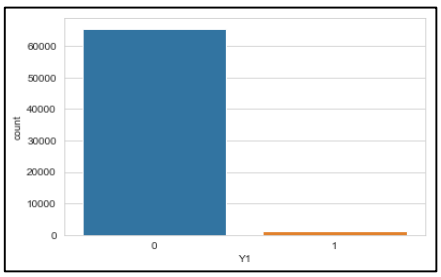

圖 9

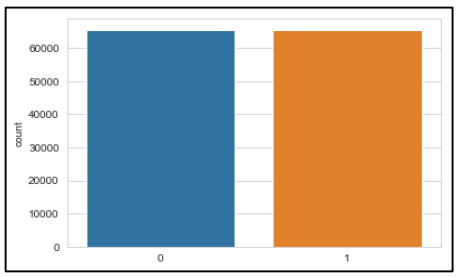

圖 10

針對 Logistic Regression 以第四種方式來運行，做 One-hot

Encoding、做 SMOTE 並調整模型參數。如圖 9，因為資料中目標變數 Y1 欄位的類別 N 與類別 Y 的分配極不平衡的關係，Logistic Regression 模型易受這種不平衡資料影響，因此對訓練資料的樣本(Training Data)做 SMOTE，取後放回創造一組新的資料，在新資料中，如圖 10，目標變數 Y1 欄位的類別 N 與類別 Y 的分配達到平衡。接著選用方法二的參數組合調整模型參數，並用最佳的參數來運行模型。做 One-hot Encoding、採用最佳的參數組合並使用 SMOTE，Train 資料集 33%的驗證資料其驗證結果的 AUC 值為 0.791。模型的上傳成績的 AUC 值為 0.8153。

針對 Logistic Regression 以第五種方式來運行，增加新的資料處理方式。新的資料處理方式並未把 AMT 系列 15 個欄位加總，每個欄位的缺失值都用中位數來填補；另外 AGE（年齡）與 IF_2ND_GEN_IND(是否為保戶二代)兩個欄位合併為新欄位，新欄位的類別包含: 低年齡者且是保戶二代、中年齡者且不是保戶二代、中高年齡者且是保戶二代共三種類別。增加新的資料處理方式後，Train 資料集 33%的驗證資料其驗證結果的 AUC 值為 0.8030。模型的上傳成績的 AUC 值為 0.8298。

五、 結果討論

最終統計，驗證資料集在選用新的資料清理方式並調參後的 XGBoost 下運行時表現最佳，實際上傳的成績則是完全未經處理的 XGBoost 表現最佳，然而平台最終還會用新的一筆資料集來測試模型作為 Private Leaderboard 的成績，因此還尚不能確定何種模型會在最後勝出。另外 Logistic Regression 不論經何種方法處理表現均差於任何一種配置的 XGBoost。

表 6

	Validation Data (AUC)	上傳成績 (AUC)
XGBoost，未處理	0.815	0.8469
XGBoost，調參	0.817	0.8407
XGBoost，調參 + 選擇特徵	0.818	0.8441
XGBoost，新的資料清理方式	0.8143	0.8440
XGBoost，新的資料清理方式+調參	0.822	0.8427
Logistic Regression，未處理	0.796	0.8321
Logistic Regression，調參 + One-hot	0.801	0.8329
Logistic Regression，調參 + One-hot + 選擇特徵	0.801	0.8329
Logistic Regression，調參 + One-hot + 重複抽樣	0.791	0.8153
Logistic Regression，新的資料清理方式	0.8030	0.8298

六、 結論

本專案的目標為透過保險客戶歷史資料來評估預測指定的客戶其在未來三個月內是否會購買重疾險保單，我們透過資料探索式分析來了解目標欄位特性、遺漏值分布、資料特徵等狀況，而針對類別型資料與缺失值資料，我們深入探討在保險情境下適合的資料清理與處理方案，最後整合六種機器學習模型方法，並透過適當的參數調整與探討，獲得現階段最佳的模型解決方案，實驗結果 XGBoost 與 AdaBoost 方法成效為最佳，而針對參賽策略的決策，我們選定了以 XGBoost 與 Logistic Regression 為基礎的模型來改良並參賽。對於本專案的未來建議，礙於時間有限的關係，最初在 4.2 節 Model Selection 的部分時，用於比較作為後續處理與上傳測試的模型均是在尚未調整參數的情況下來進行比較，這也代表了表 6 中這些模型的成績並非模型本身的最佳表現，而如果在每個模型都未發揮其最佳水準的比較基礎下，有可能因而錯失了更具潛力的模型作為上傳測試與強化的機會。因此，如果未來能在 4.2 節 Model Selection 的部分進行改善，將可能有機會訓練出表現更加的模型。

七、 附件

7.1 保險公司的顧客資料來源分析

一般一個投保的行為包含了，要保人、被保險人、受益人、保險人(保險公司)4 個角色，除了保險人之外的要保人、被保險人、受益人為保險公司的顧客，在投保時會進行相關資料的填寫。填寫的資料主要包括 3 項:1.基本資料 2. 健康告知書 3.業務員詢問的生調項目。其中基本資料為要保人、被保險人、受益人皆須填寫的項目，健康告知書只有被保險人需填寫，生調項目則一般詢問要保人、被保險人收入、婚姻狀況等資訊。因此這個比賽的資料來源應該就出於這 3 的地方。而這些欄位若有遺漏值可能是由於該身分的人不需要填寫該欄位資訊所致(e.g 受益人不會被詢問生調項目，因此其年收入欄位可能就是遺漏值)。

7.2 特徵分析（僅針對 Train 資料集）

一般保險公司的客戶名單來源，應該包含了要保人、被保險人、受益人(詳附錄 7.1)。首先，觀察後發現「APC_1ST_AGE 首次擔任要保人年齡」、「RFM_M_LEVEL 曾投保主約件數」、「REBUY_TIMES_CNT 再購次數」、「TERMINATION_RATE 曾解約保單張數佔曾投保保單張數佔率」、「APC_1ST_YEARDIF 首次成為要保人距今間隔時間」有完全相同的 43282 筆遺失值(同樣的幾名客戶)，從這些欄位共同的特性判斷，這些遺失值可能屬於「非要保人」的顧客，因此無法填答這些欄位的問題。另外，可以發現這些欄位是遺失值的客戶資料，在 A、B 通路投保的新契約數皆為 0，目前是否持有壽險保單相關欄位皆顯示為否，服務人員、業務人員的數量為 0，是否催繳、是否有投保附約也都顯示為否，由於這些欄位的結果皆符合「非要保人」身分應該有的結果，因此我們推測這 43282 人即為「非要保人」身分。而 LEVEL 往來關係等級，由於一般和保險公司直接接觸的是要保人，因此 LEVEL 的紀錄也應該會以要保人為主,所以在本段提及的這些欄位有遺漏值者,也就是我們推測的「非要保人」,應該也傾向 LEVEL 屬於遺漏值，觀察後發現 LEVEL 中有 43305 筆遺漏值，其中包含的 43281 筆對應到本段提及的欄位也為遺漏值，僅 1 人沒對應到(以要保人紀錄為主，仍可能有非要保人)；RFM_R 則是非要保人無法填寫的項目，RFM_R 中有 43294 個遺失

值，其中有 43282 筆為對應到我們所說的「非要保人」資料，剩餘的 12 筆可能為其他原因所致的遺漏值。

其次，表 7、8 為 MARRIAGE_CD、EDUCATION_CD、ANNUAL_INCOME_AMT 三欄位共同是遺漏值的情況下，INSD_1ST_AGE、APC_1ST_AGE 分別的分配表，由於 MARRIAGE_CD、EDUCATION_CD、ANNUAL_INCOME_AMT 這三項都是業務員生調的項目，而不出現在基本資料或健康報告書(詳附錄 7.1)，而這些項目只有被保險人和要保人需要回答，因此若純受益人(非要保且非被保)身分的人存在則應包含於上述 3 欄位共同遺失值中，但根據表 7、8 的結果顯示，這三欄位都為遺漏值的客戶，仍大多有擔任要保人或被保人，意思是最多也只有 14 名(NaN)可能是純受益人，由於數量極少所以可以忽視，視為沒有純受益人。因此若完全不考慮受益人身分存不存在，第一段所提到的非要保人，就可以解釋成是純被保人。同理，也可以從「IF_ADD_INSD_IND 是否投保附約(被保)」、「INSD_1ST_AGE 首次擔任被保人年齡(級距)」、「INSD_LAST_YEARDIF_CNT 最近一次被保人身份投保距今間隔時間(年)」這三項欄位的缺失值情況與不存在純受益人的事實去推論有純要保人的存在，原理與推論存在純被保人的方法相同。推測這個資料集中僅有 3 種身分:純要保人(非被保)、純被保人(非要保)、同時為要保人及被保險人。(其中這些身分的人可能是受益人，也可能不是)。

表 7:「INSD_1ST_AGE」欄位次數

INSD_1ST_AGE (首次擔任被保人年齡)	資料筆數
NaN	14
3	1939
1	1961
4	1660
2	1540

表 8:「APC_1ST_AGE」欄位次數表

APC_1ST_AGE (首次擔任要保人年齡)	資料筆數
NaN	3620
3	1388
1	791
4	774
2	541

最後，EDUCATION 欄位分為四個類別。由圖 11 各教育程度類別之年齡分配可觀察出 4 是數量最少的類別，因此推測為研究所，而 1 有很多低年齡者，推測為國中以下，2、3 由順序上來看合理推估為高中職、大專院校(且目前人口統計資料以大專院校人口最多)，由於

最低的年齡在研究所仍存在，因此最低年齡的上限(中年齡的下限)應該在 2x 歲左右。

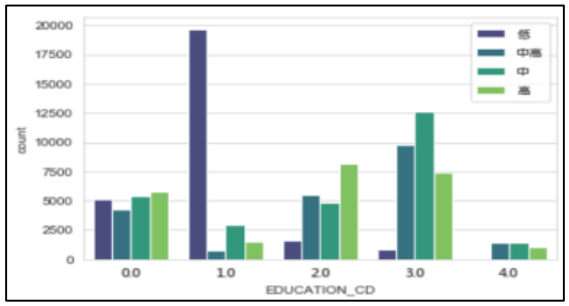

圖 11 ： 各教育程度類別之年齡分配

觀察圖 12 各教育類別之年齡的 Y1 欄位之 Y/N 分布中，教育程度為 3 且是中或中高年齡有著特別高的 Y 數量(購買重疾險數量)(0 表示教育程度為 NaN)，從比例來看(參考圖 13)：教育程度為 3 且是中或中高年齡的 Y1 也有高比例，教育程度為 2 且是中或中高年齡也顯示有高比例。

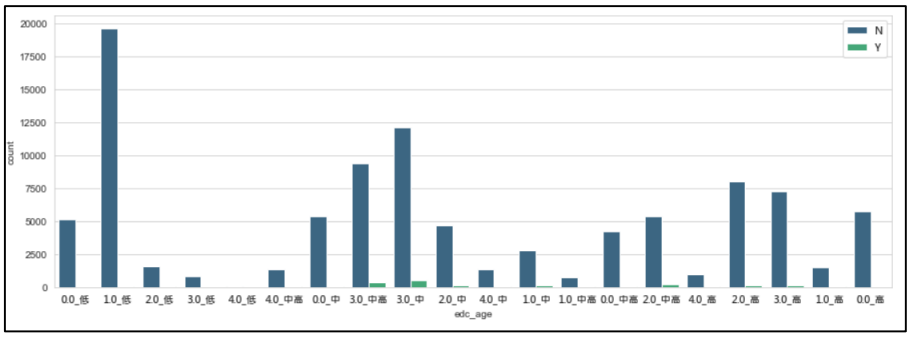

圖 12：教育類別年齡的 Y1 欄位之 Y/N 分布(X 軸為:教育程度_年齡)

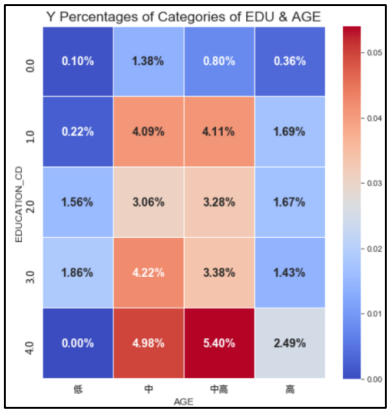

圖 13：教育程度類別與年齡的 Y1 欄位之 Y/N 分布比例