國泰大數據競賽　Dcslab_分析說明書

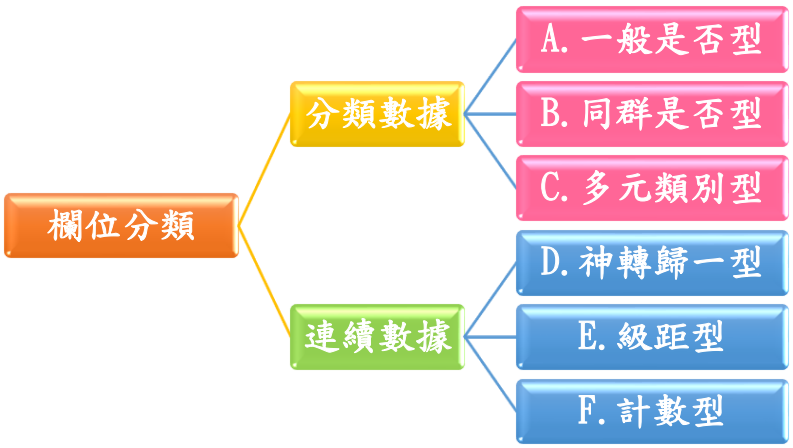
第一部分　資料預處理與特徵選擇

一、資料預處理流程圖：

二、資料預處理流程摘要：

由於各欄位資料的處理方式不盡相同，但我們仍可分成六大類來處理。再來以客戶購買行為(詢問專家經驗)搭配描述性統計的方式來篩選欄位，這步驟主要以各級距或各類別的「購買率」為指標來判斷。第三步會根據缺失值的購買率及其分別在訓練集與測試集的比例分佈，來決定缺失值的處理方式。異常值的部分會視情況移除或設一個天花板值。第四步變數變換的方法中，我們盡可能不增加各欄位的維度來處理，也就是用連續數據與分類數據混成的處理方式，將各欄位都設計成仍具有分類數據特性的連續數據，各欄位用一個維度就能表達行為，並均以購買率做為參考。在某些同群且購買行為統計接近的欄位，我們甚至將其合併以降低維度。最後即形成第五步訓練模型用的特徵向量。

三、欄位分類：



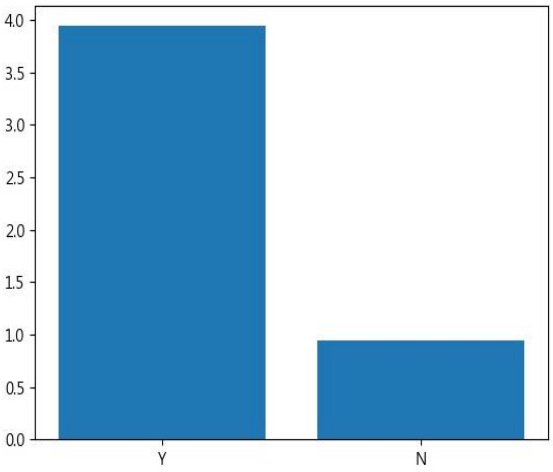
一般來說資料有「分類數據」與「連續數據」兩大類，而仔細觀察本次競賽中可用於當訓練資料的 130 個欄位中(扣除 CUS_ID 客戶流水編號與 Y1 預測目標二欄)，仍可再分成六大類，如下圖：

四、各類別欄位處理說明

A. 分類數據--「一般是否型」:

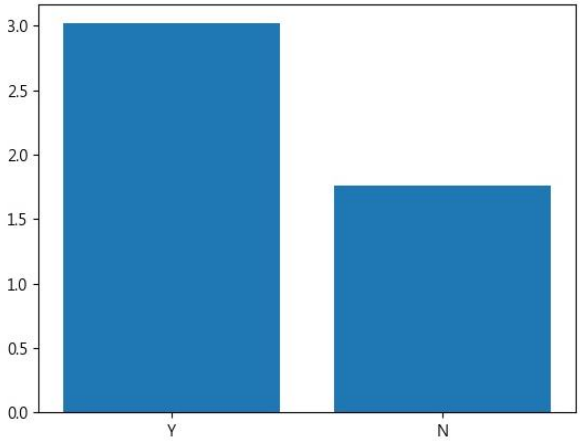
欄位編號	欄位英文名稱	NA train	NA test	欄位中文名稱	特徵評分	轉換處理方式概述
8	LAST_A_CCONTACT_DT	0	0	近三年是否有與 A 通路接觸	3	正常 01
10	LAST_A_ISSUE_DT	0	0	近三年是否有透過 A 通路投保新契約	3	正常 01
12	LAST_B_ISSUE_DT	0	0	近三年是否有透過 B 通路投保新契約	0	刪掉
20	IF_2ND_GEN_IND	0	0	是否為保戶二代	1	正常 01
49	IF_ADD_IND	0	0	是否投保附約(要保)	3	正常 01
55	L1YR_PAYMENT_REMINDER_IND	0	0	近一年是否曾催繳	1	正常 01
56	L1YR_LAPSE_IND	0	0	近一年是否曾停效	0	刪掉
57	LAST_B_CONTACT_DT	0	0	近三年是否有與 B 通路接觸	2	正常 01
61	LAST_C_DT	0	0	近三年是否有到 C 通路申辦服務	2	正常 01
66	IF_S_REAL_IND	0	0	是否投保 S 險(Y/N)(附約)	1	正常 01
67	IF_Y_REAL_IND	0	0	是否投保 Y 險(Y/N)(附約)	0	刪掉
98	IF_HOUSEHOLD_CLAIM_IND	0	0	家人是否曾申請理賠(Y/N)	1	正常 01
122	IF_ADD_INSD_IND	0.171	0.0001	是否投保附約(被保)	2	正常 01，NA 設 mean 值

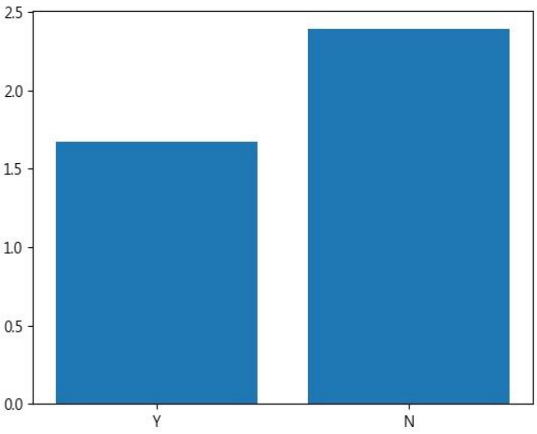
上表為「一般是否型」的欄位清單。其中 NA train 及 NA test 分別為該欄位的訓練集與測試集之 NA 比例統計。關於「特徵評分」，是基於該欄中各類別之購買率比較結果，特徵評分程度為：「(0 不顯著)、(1 一般般)、(2 具特徵)、(3 顯著)」，舉例說明如下：



左圖為 LAST_A_CCONTACT_DT 的 Y 和 N 的購買率直方圖(沒有缺失值)，其中 Y 的購買率約為 4%比 N 的 1%高出約 3%，並且 Y 與 N 的客戶量分別為 35405 人與 64595 人。也就是說與 A 通路有接觸的客戶有 35405 人，並且購買率明顯多於與 A 通路沒接觸的人，即此欄位具明顯特徵，評分為 3 分。

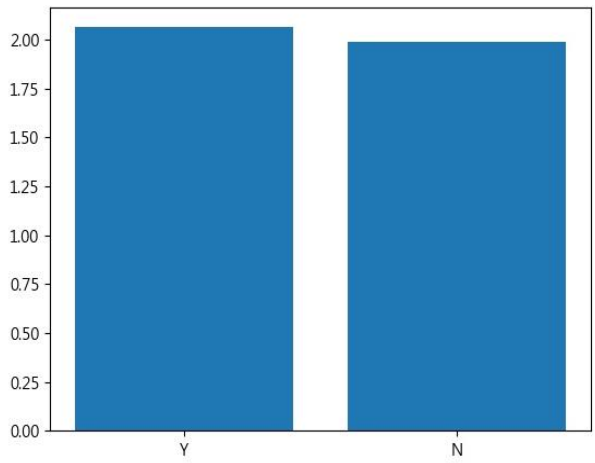
右圖 LAST_B_CONTACT_DT 的 Y 和 N 購買率圖，其中 Y 購買率約為 3%比 N 的 1.6%高出約 1.4%，並且 Y 與 N 的客戶量分別為 19188 人與 80812 人。也就是說，近三年有接觸 B 通路的客戶購買重疾險的比例多出了 1.4%，具備特徵，這裡評分為 2 分。





左圖 IF_2ND_GEN_IND 的 Y 和 N 購買率圖，其中 N 購買率約為 2.4%比 Y 的 1.6%高出約 0.8%，並且 Y 與 N 的客戶量分別為 54315 人與 45685 人。也就是說，非保戶二代的客戶，比是保戶二代的客戶購買重疾險的比例只多出 0.8%，特徵一般，這裡評分為 1 分。

右圖 IF_Y_REAL_IND 的 Y 和 N 購買率圖，其中 Y 購買率約為 2.05%和 N 的 1.95%僅高出約 0.1%，並且 Y 與 N 的客戶量分別為 17888 人與 82112 人。也就是說，不管有沒有投保 Y 險，購買率沒有顯著差異，因些不具備特徵，這裡評分為 0 分。



轉換的部分，因為各欄只有 Y 與 N 二元類別，所以均轉成 1 和 0。此類唯欄位 122 IF_ADD_INSD_IND 雖有缺失值，然而在訓練集比例卻不高，且在測試集比例更低，所以此欄位缺失值取 mean 值。

B. 分類數據--「同群是否型」：

這部分欄位最多，各系列處理方式分述如下：

欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	轉換處理方式概述
69	IM_IS_A_IND	0	0	是否持有特定商品_A(Y/N)	0	刪除
70	IM_IS_B_IND	0	0	是否持有特定商品_B(Y/N)	2	正常 01
71	IM_IS_C_IND	0	0	是否持有特定商品_C(Y/N)	1	
72	IM_IS_D_IND	0	0	是否持有特定商品_D(Y/N)	1	
74	X_A_IND	0.00038	0.0001	是否申辦 A 服務(Y/N)	1	正常 01，NA 取 mean
75	X_B_IND	0.00038	0.0001	是否申辦 B 服務(Y/N)	3	
76	X_C_IND	0.00038	0.0001	是否申辦 C 服務(Y/N)	2	
77	X_D_IND	0.00038	0.0001	是否申辦 D 服務(Y/N)	1	
78	X_E_IND	0.00038	0.0001	是否申辦 E 服務(Y/N)	1	
79	X_F_IND	0.00038	0.0001	是否申辦 F 服務(Y/N)	1	
80	X_G_IND	0.00038	0.0001	是否申辦 G 服務(Y/N)	1	
81	X_H_IND	0.00038	0.0001	是否申辦 H 服務(Y/N)	3	

以上的兩群處理與 A 型別類同。

58	A_IND	0.79871	0.790807	是否訂閱 A 電子報(Y/N)	2	(NA：0， N:0.8， Y:1)
59	B_IND	0.79871	0.790807	是否訂閱 B 電子報(Y/N)	2	
60	C_IND	0.79871	0.790807	是否訂閱 C 電子報(Y/N)	2	
125	FINANCETOOLS_A	62.641	0.6235	理財工具_A	1	(NA：-1， N:0， Y:1)
127	FINANCETOOLS_C	62.641	0.6235	理財工具_C	1	(NA：-1， N:0， Y:1)
126	FINANCETOOLS_B	62.641	0.6235	理財工具_B	1	(NA：-1， N:0， Y:1, 1.26, 1.442)
128	FINANCETOOLS_D	62.641	0.6235	理財工具_D	1	
130	FINANCETOOLS_F	62.641	0.6235	理財工具_F	1	
129	FINANCETOOLS_E	62.641	0.6235	理財工具_E	2	(NA：-1， N:0， Y:1, 1.1412)
131	FINANCETOOLS_G	62.641	0.6235	理財工具_G	1	

其中是否訂閱電子報的統計如下：

Name	N_count	Y_count	NA_count	N_Y_rate	Y_Y_rate	NA_Y_rate
A_IND	12979	7150	79871	0.042	0.0435	0.0143
B_IND	16022	4107	79871	0.0407	0.05	0.0143
C_IND	19737	839	79871	0.0425	0.046	0.0143

此欄位缺失值佔比在訓練集與測試集均超過 79%。而在訓練集中，NA 購買比例均只有 1.43%，且無論有無訂閱 ABC 三電子報的客戶，購買重疾險的比例均高達 4~5%，因此我們決定將 NA 值視為一個類別，賦值方式為：(NA：0， N:0.8， Y:1)。

再來是是否使用理財工具 A~G：

Name	N_count	Y_count	NA_count	N_Y_rate	Y_Y_rate	NA_Y_rate
FINANCETOOLS_A	10903	26456	62641	0.024397	0.030617	0.014751
FINANCETOOLS_C	24837	12522	62641	0.028224	0.029947	0.014751
FINANCETOOLS_B	33826	3533	62641	0.028735	0.029437	0.014751
FINANCETOOLS_D	34308	3051	62641	0.028711	0.029826	0.014751
FINANCETOOLS_F	36161	1198	62641	0.028705	0.03172	0.014751
FINANCETOOLS_E	37114	245	62641	0.028884	0.016327	0.014751
FINANCETOOLS_G	34101	3258	62641	0.029383	0.022713	0.014751

觀察 NA 值也不少，訓練集與測試集均超過 62%，並且 NA_Y_rate 均比 N_Y_rate 及 Y_Y_rate 低了一半，因此決定將 NA 視為一類，賦值為-1，然後將 N 設為 0、Y 設為 1。

此外，經觀察購買率後，BDF 的購買行為類似，所以合併成一個欄位，並且 Y 的賦值方式為：若 BDF 均有者為 3，只有 BD 或 DF 或 BF 者為 2，只有 B 或 D 或 F 者為 1，然後因為合併後最大值為 3，所以再將 Y 值開三次方根。(註：原因是會用到 CNN 模型，我們希望將最後特徵向量的值 normailze 到-1~1.5 之間)。而 EG 兩欄的合併方式類同 BDF。

欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	轉換處理方式概述
27	IF_ISSUE_A_IND	0	0	目前是否壽險保單持有有效類別_A(主約)	1	N:0 Y:1, 1.189 1.316, 1.4142
31	IF_ISSUE_E_IND	0	0	目前是否壽險保單持有有效類別_E(主約)	1	
32	IF_ISSUE_F_IND	0	0	目前是否壽險保單持有有效類別_F(主約)	1	
37	IF_ISSUE_K_IND	0	0	目前是否壽險保單持有有效類別_K(主約)	1	
47	IF_ADD_G_IND	0	0	目前是否壽險保單持有有效類別_G(附約)	1	
28	IF_ISSUE_B_IND	0	0	目前是否壽險保單持有有效類別_B(主約)	1	N:0， Y:1, 1.26, 1.442
33	IF_ISSUE_G_IND	0	0	目前是否壽險保單持有有效類別_G(主約)	2	
36	IF_ISSUE_J_IND	0	0	目前是否壽險保單持有有效類別_J(主約)	3	

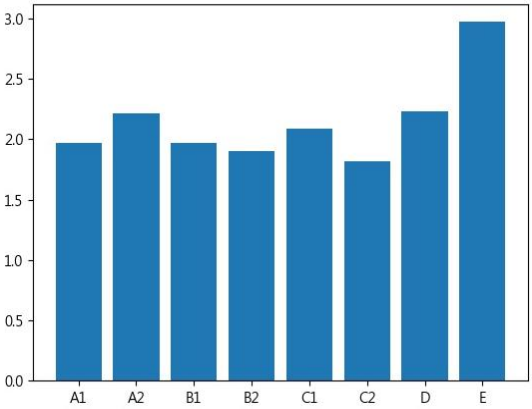
29	IF_ISSUE_C_IND	0	0	目前是否壽險保單持有有效類別_C(主約)	2	N:0， Y:1, 1.1412
30	IF_ISSUE_D_IND	0	0	目前是否壽險保單持有有效類別_D(主約)	2	
34	IF_ISSUE_H_IND	0	0	目前是否壽險保單持有有效類別_H(主約)	1	N:0， Y:1, 1.26, 1.442
35	IF_ISSUE_I_IND	0	0	目前是否壽險保單持有有效類別_I(主約)	3	
38	IF_ISSUE_L_IND	0	0	目前是否壽險保單持有有效類別_L(主約)	1	
40	IF_ISSUE_N_IND	0	0	目前是否壽險保單持有有效類別_N(主約)	2	N:0， Y:1, 1.26, 1.442
42	IF_ISSUE_P_IND	0	0	目前是否壽險保單持有有效類別_P(主約)	2	
44	IF_ADD_F_IND	0	0	目前是否壽險保單持有有效類別_F(附約)	2	
43	IF_ISSUE_Q_IND	0	0	目前是否壽險保單持有有效類別_Q(主約)	3	N:0， Y:1, 1.26, 1.442
45	IF_ADD_L_IND	0	0	目前是否壽險保單持有有效類別_L(附約)	3	
46	IF_ADD_Q_IND	0	0	目前是否壽險保單持有有效類別_Q(附約)	3	
48	IF_ADD_R_IND	0	0	目前是否壽險保單持有有效類別_R(附約)	3	正常 0 1
39	IF_ISSUE_M_IND	0	0	目前是否壽險保單持有有效類別_M(主約)	0	刪掉
41	IF_ISSUE_O_IND	0	0	目前是否壽險保單持有有效類別_O(主約)	0	刪掉

100	IF_ISSUE_INSD_A_IND	0.201	0.199	目前是否壽險保單被保有效類別_A(主約)	1	N:0， Y:1, 1.1412
103	IF_ISSUE_INSD_D_IND	0.201	0.199	目前是否壽險保單被保有效類別_D(主約)	1	
101	IF_ISSUE_INSD_B_IND	0.201	0.199	目前是否壽險保單被保有效類別_B(主約)	1	N:0， Y:1, 1.26, 1.442
102	IF_ISSUE_INSD_C_IND	0.201	0.199	目前是否壽險保單被保有效類別_C(主約)	2	
106	IF_ISSUE_INSD_G_IND	0.201	0.199	目前是否壽險保單被保有效類別_G(主約)	2	
105	IF_ISSUE_INSD_F_IND	0.201	0.199	目前是否壽險保單被保有效類別_F(主約)	1	N:0 Y:1, 1.189 1.316, 1.4142
112	IF_ISSUE_INSD_M_IND	0.201	0.199	目前是否壽險保單被保有效類別_M(主約)	1	
113	IF_ISSUE_INSD_N_IND	0.201	0.199	目前是否壽險保單被保有效類別_N(主約)	1	
114	IF_ISSUE_INSD_O_IND	0.201	0.199	目前是否壽險保單被保有效類別_O(主約)	1	
107	IF_ISSUE_INSD_H_IND	0.201	0.199	目前是否壽險保單被保有效類別_H(主約)	1	N:0， Y:1, 1.1412
108	IF_ISSUE_INSD_I_IND	0.201	0.199	目前是否壽險保單被保有效類別_I(主約)	3	
109	IF_ISSUE_INSD_J_IND	0.201	0.199	目前是否壽險保單被保有效類別_J(主約)	2	N:0， Y:1, 1.26, 1.442
110	IF_ISSUE_INSD_K_IND	0.201	0.199	目前是否壽險保單被保有效類別_K(主約)	1	
111	IF_ISSUE_INSD_L_IND	0.201	0.199	目前是否壽險保單被保有效類別_L(主約)	1	
118	IF_ADD_INSD_L_IND	0.518	0.516	目前是否壽險保單被保有效類別_L(附約)	2	N:0， Y:1, 1.26, 1.442
119	IF_ADD_INSD_Q_IND	0.518	0.516	目前是否壽險保單被保有效類別_Q(附約)	2	
120	IF_ADD_INSD_G_IND	0.518	0.516	目前是否壽險保單被保有效類別_G(附約)	1	
115	IF_ISSUE_INSD_P_IND	0.518	0.516	目前是否壽險保單被保有效類別_P(主約)	2	正常 01
116	IF_ISSUE_INSD_Q_IND	0.518	0.516	目前是否壽險保單被保有效類別_Q(主約)	2	正常 01
117	IF_ADD_INSD_F_IND	0.518	0.516	目前是否壽險保單被保有效類別_F(附約)	1	正常 01
121	IF_ADD_INSD_R_IND	0.518	0.516	目前是否壽險保單被保有效類別_R(附約)	1	正常 01
104	IF_ISSUE_INSD_E_IND	0.201	0.199	目前是否壽險保單被保有效類別_E(主約)	0	

以上 44 個欄位處理方法&合併與「是否使用理財工具 A~G」類同。

C. 分類數據--「多元類別型」:

欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	轉換處理方式概述
2	GENDER(2 類)	0.0068	0.0067	性別	2	正常 01
4	CHARGE_CITY_CD (8 類)	0	0	收費地址_縣市	2	B1 此類賦值 1 其它依購買率設等差
5	CONTACT_CITY_CD(8 類)	0	0	聯絡地址_縣市	1	刪掉
7	MARRIAGE_CD(3 類)	0.0795	0.0765	婚姻狀況	2	NA=-1，其它等差
15	OCCUPATION_CLASS_CD (7 類)	0.0396	0.0328	客戶職業類別(各類別)對核保風險程度	2	變 conti (NA 取 1.2)
124	CUST_9_SEGMENTS_CD	0	0	九大客群(只有 8 類)	3	

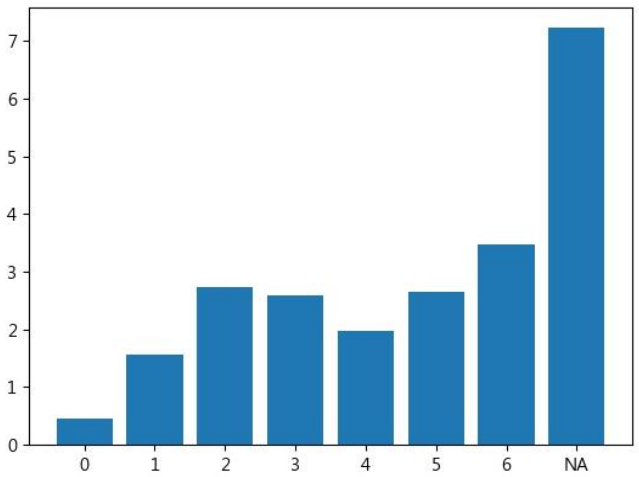


A1	27066
B1	18680
A2	14489
C2	14454
B2	13027
C1	7961
D	3852
E	471

左圖為 CONTACT_CITY_CD 的購買率統計，會篩掉這個欄位的原因，主要是因為各類別並沒有顯著的購買率差異。再者購買率最高的 E 類，其客戶人數也只有 471 位。因此特徵向量將不考慮此欄位。

```
# OCCUPATION_CLASS_CD
c15=np.zeros(data.shape[0])
c15[data[:,14]=='1']=0.4
c15[data[:,14]=='2']=0.7
c15[data[:,14]=='3']=0.7
c15[data[:,14]=='4']=0.5
c15[data[:,14]=='5']=0.7
c15[data[:,14]=='6']=0.9
c15[data[:,14]=='NA']=1.2
```

1.0	74372
2.0	14517
3.0	3827
4.0	2025
0.0	871
5.0	226
6.0	202

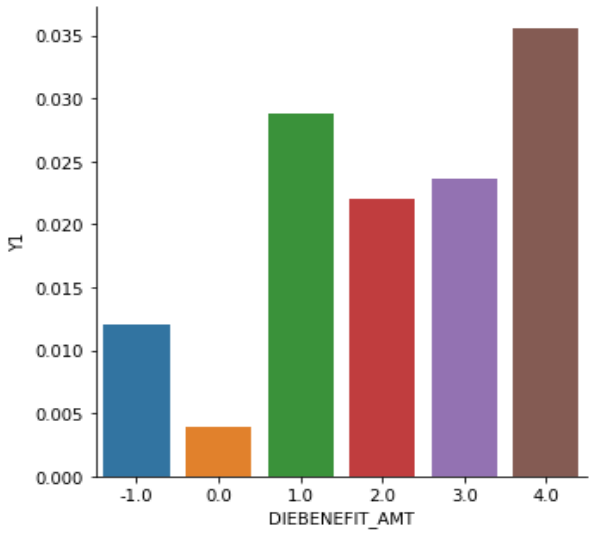


接下來試舉一個我們將多元類別的欄值，轉成連續數據的處理方法：以上右圖是 OCCUPATION_CLASS_CD 的各類別購買率直方圖，中間是各類別人數統計，左圖是我們參考購買率高低，將 8 種類別(含 NA)從分類數據轉連續數據的例子。如此做的概念是：雖然職業有 7 種，連同購買率最高的 NA 視為另一種，總共 8 種，但其實都同屬一個欄位，用 onehot encoding 的話，反而會讓關係變太各自獨立 (independent)，所以參考購買率高低，將這類欄位轉成連續數據。

D. 連續數據--「神轉歸一型」:

欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	轉換處理方式概述
21	APC_1ST_YEARDIF	0.4328	0.4281	首次成為要保人距今間隔時間	2	
50	ANNUAL_PREMIUM_AMT	0.6244	0.6182	年繳化保費	1	
54	ANNUAL_INCOME_AMT	0.3920	0.3827	年收入	1	
63	BANK_NUMBER_CNT	0	0	銀行往來家數(轉帳)	2	
65	BMI	0.1664	0.1579	BMI 值	1	NA 取 mean
83	DIEBENEFIT_AMT	0.2754	0.2764	當年度保障_一般身故	2	
84	DIEACCIDENT_AMT	0.2754	0.2764	當年度保障_意外身故	2	
85	POLICY_VALUE_AMT	0.2754	0.2764	當年度保障_保單價值	2	
86	ANNUITY_AMT	0.2754	0.2764	當年度保障_年金	3	
87	EXPIRATION_AMT	0.2754	0.2764	當年度保障_滿期金	3	
88	ACCIDENT_HOSPITAL_REC_AMT	0.2754	0.2764	當年度保障_意外醫療住院日額保險金	1	
89	DISEASES_HOSPITAL_REC_AMT	0.2754	0.2764	當年度保障_疾病醫療住院日額保險金	1	
90	OUTPATIENT_SURGERY	0.2754	0.2764	當年度保障_手術醫療門診手術	1	
91	INPATIENT_SURGERY	0.2754	0.2764	當年度保障_手術醫療住院手術	1	
92	PAY_LIMIT_MED_MISC_AMT	0.2754	0.2764	當年度保障_實支實付醫療及雜費限額	3	
93	FIRST_CANCER_AMT	0.2754	0.2764	當年度保障_癌症醫療初次罹癌給付	2	
94	ILL_ACCELERATION_AMT	0.2754	0.2764	當年度保障_重大疾病提前給付	3	
95	ILL_ADDITIONAL_AMT	0.2754	0.2764	當年度保障_重大疾病額外給付	3	
96	LONG_TERM_CARE_AMT	0.2754	0.2764	當年度保障_長看照顧保險金	3	
97	MONTHLY_CARE_AMT	0.2754	0.2764	當年度保障_傷殘給付每月生活照護保險金	2	
99	LIFE_INSD_CNT	0	0	目前主約被保有效件數(件)	2	
123	L1YR_GROSS_PRE_AMT	0	0	近一年實繳保費	3	

觀察以上欄位，有 NA 值者，其佔比至少都有 16%以上。而數值為零者，其佔比大約在 13%~72%之間。其它非 NA 且數值大於零者，為連續數據我們用四分位數來分群比較分析其購買率。本類別欄位，處理方式類似，經比較數量及購買率後，NA 值我們設置為-1，原為零者就為 0，連續數據部分則安排在 0.5~1.5 之間。另外針對數量級很小例如 10^{-7}、10^{-6} 與 10^{-5} 同時都各有上萬筆者，我們有取 log 再 rescale 到 0.5~1.5 之間的處理。

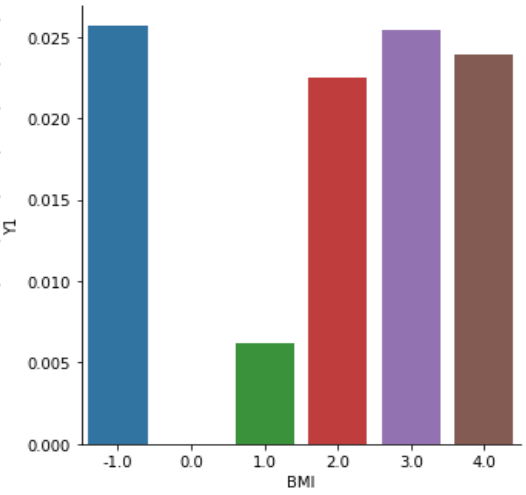


-1	27540
0	13734
1	14683
2	14681
3	14723
4	14639

左圖以 DIEBENEFIT_AMT 為例，-1 為 NA 的數量有 27540 個，購買率約 1.2%。0 為零值的數量有 13734 個，購買率為 0.4%。其它大於零者有 58726 個，然後劃分四分位數做購買率比較，發現均有 2.2%以上。因此將欄位的值設計成(NA:-1、零:0、大於零:0.5~1.5)，讓這類欄位同時有連續&分類數據的特性。

右圖是 BMI 是本類別欄位稍特殊的例子。-1 為 NA 且數量有 16645 筆，購買率約 2.6%與第三個四分位距的購買相近。然而零值者數量只有 63 個。因此，此欄我們決定將 NA 值取平均，其它數值維持原本的大小，所以此欄位數值介於 0~1 之間。

-1	16645
0	63
1	22761
2	26672
3	15679
4	18180



E. 連續數據--「級距型」:

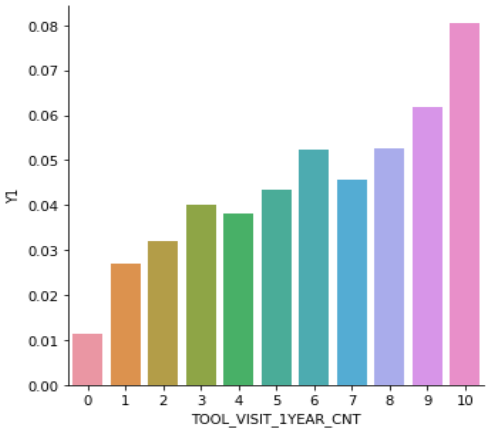
欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	NA 處理方式概述
3	AGE	0	0	年齡(年)(級距)	2	轉成等差
6	EDUCATION_CD	0.20562	0.202	教育程度/學歷	3	NA 取 0
18	APC_1ST_AGE	0.43282	0.4281	首次擔任要保人年齡(級距)	1	NA 取-1
19	INSD_1ST_AGE	0.00171	0.0001	首次擔任被保人年齡(級距)	2	NA 取 mean
22	RFM_R	0.43294	0.4282	上次要保人身份投保距今間隔時間	3	NA 取-1
23	REBUY_TIMES_CNT	0.43282	0.4281	再購次數(級距)	2	NA 取-1
24	LEVEL	0.43305	0.4281	往來關係等級(1~5 級)	3	NA 取-1
25	RFM_M_LEVEL	0.43282	0.4281	曾投保主約件數(等級)	3	NA 取-1
26	LIFE_CNT	0	0	目前主約持有有效件數(件)(級距)	3	轉成等差

此類型的處理方向，就是將各欄位自己的級距轉成大約在 0~1.5 之間的等差級數。原因是這些欄的級距在本質上並非是 Independent 的，若用 onehot encoding 的處理方式，反而有違其本質。至於 NA 的部分，其分析處理流程與其它類別的欄位一樣。

F. 連續數據--「計數型」:

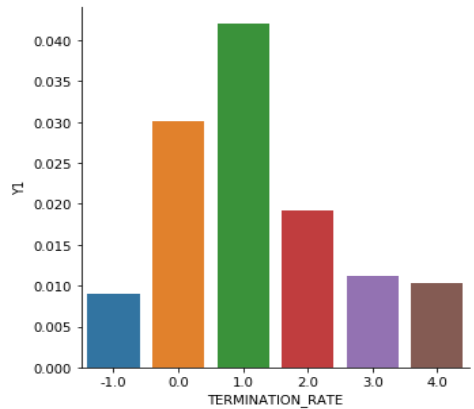
欄位編號	欄位英文名稱	NA-train	NA-test	欄位中文名稱	特徵評分	轉換處理方式概述
9	L1YR_A_ISSUE_CNT	0	0	近一年透過 A 通路投保新契約次數	2	變 01
13	CHANNEL_A_POL_CNT	0	0	透過 A 通路投保新契約件數	3	變 01
17	INSD_CNT	0	0	對應的被保人數	1	變 01
51	AG_CNT	0	0	曾經經手過的業務員人數	2	變 01
52	AG_NOW_CNT	0	0	曾經經手過且目前在職業務員人數	3	變 01
53	CLC_CUR_NUM	0	0	目前服務人員人數	3	變 01
16	APC_CNT	0	0	對應的要保人數	2	Normalize
68	IM_CNT	0	0	特定商品持有類別數	2	Normalize
82	TOOL_VISIT_1YEAR_CNT	0	0	近一年業務員管理工具拜訪次數	3	Normalize
62	L1YR_C_CNT	0.8794	0.881	近一年到 C 通路申辦服務次數	1	NA 轉-1
73	TERMINATION_RATE	0.4328	0.4281	曾解約保單數佔曾投保保單數佔率	1	NA 轉 1.2
11	L1YR_B_ISSUE_CNT	0	0	近一年透過 B 通路投保新契約次數	0	刪掉
14	CHANNEL_B_POL_CNT	0	0	透過 B 通路投保新契約件數	0	刪掉
64	INSD_LAST_YEARDIF_CNT	0.0017	0.0001	最近一次被保人身份投保距今時間	0	刪掉

經觀察後，欄位編號 9、13、17、51、52、53，數值為零者大約有 50%~90%之間，然後數值非零者的購買率大於零者約有 2%~7%，已具顯著差異性，所以這些欄位以最簡單的 0 1 二元方式轉換。



0	68792
1	10256
2	5457
3	3566
4	2443
5	1843
6	1340
7	1094
7<x<=10	2126
10<x<=15	1603
15<x<=176	1480

左圖是重新 normailze 的例子 TOOL_VISIT_1YEAR_CNT 中，可以發現拜訪次數低有 98.5%都集中在 15 次以內，0 次者高達 68792 個，再來發現 outlier 值有到 176 次，所以天花板值設在 15 次，然後再整體 scale 到 0~1.5 之間。



-1	43282
0	44093
1	3739
2	4005
3	1076
4	3805

左圖最後再試舉另一種 NA 可能的處理方式。首先是 NA 與零值者的數量相當，再來是 NA 的購買率比零值低超過 2%，然後比第 4 群的購買率還要再低些。再加上我們目前最好的模型是 XGBoost，所以我們認為將 NA 值設成 1.2 來放到數列的最右側，可以讓 XG 訓練時可以有較佳的基尼值來決定節點。

第二部分 模型選擇與驗證成效說明

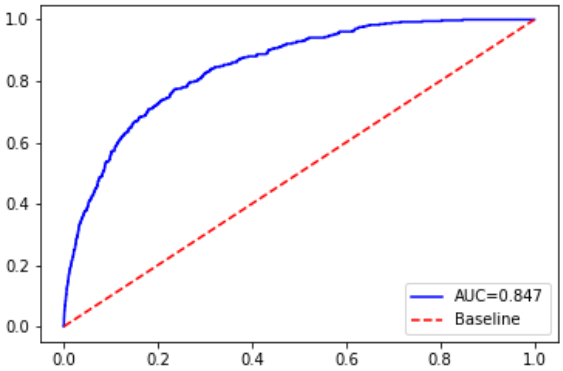
一、模型驗證流程說明：

經第一部分篩選與相關欄位合併後，預計用於模型訓練的欄位有 94 個，也就是說特徵向量的維度為 94。接下來機器學習(ML)和深度學習(DL)模型各挑若干個來進行訓練、測試與比較，然後各挑一個最佳的模型進行下一步。第三步則是測試較佳的訓練集與驗證集(validation set)的比例、進行 k-fold 交叉驗證。此外，由於本次競賽資料中正向樣本(Y1=1 者)只有 2000 筆，只佔了 10 萬筆中的 2%，屬非常不平衡的數據分析，所以我們也測試了過採樣的手法。第四步進行參數精調來比較 ML 和 DL 的哪個模型最佳。最後我們發現 XGB 是我們這次最佳的預測模型。

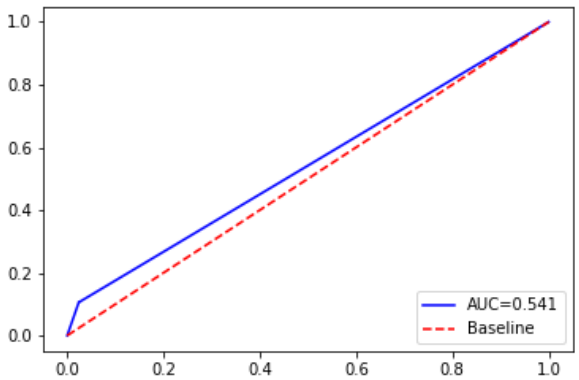
二、模型驗證流程說明：

我們選擇了較為知名的幾種模型，其中 ML 跟 DL 都挑了幾種常見的方法，並將每個模型之 ROC curve 繪製出來，比較各個模型的 AUC 分數。首先是 ML 的部分，如下圖：

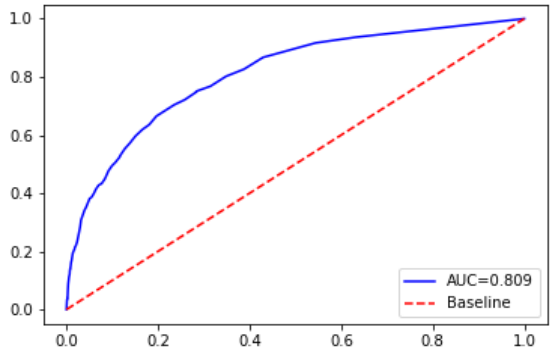
XGBoost(AUC = 0.847)



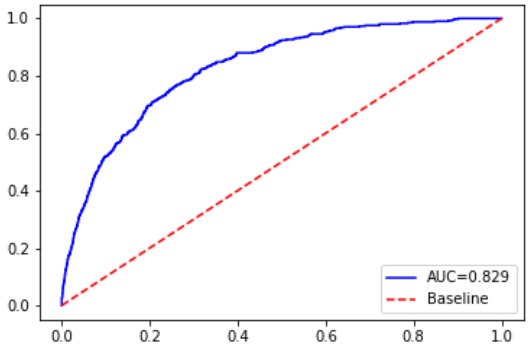
Decision Tree(AUC = 0.541)



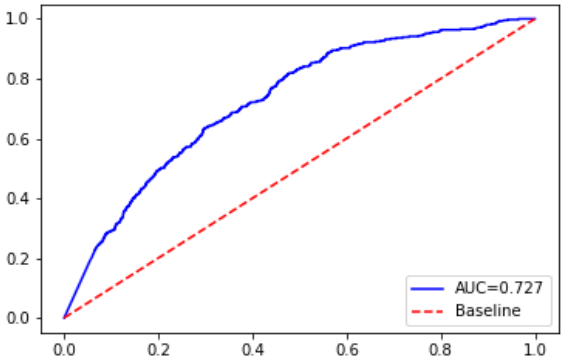
Random Forest(AUC = 0.809)



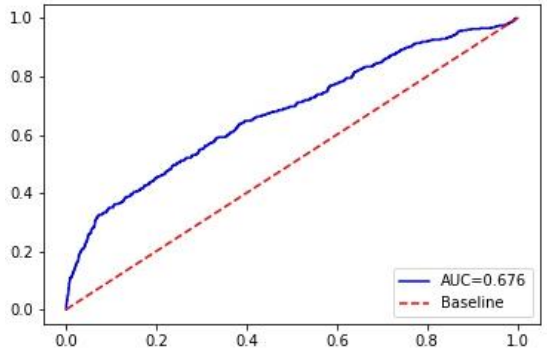
Logistic Regression(AUC = 0.829)



GaussianNB(AUC = 0.727)

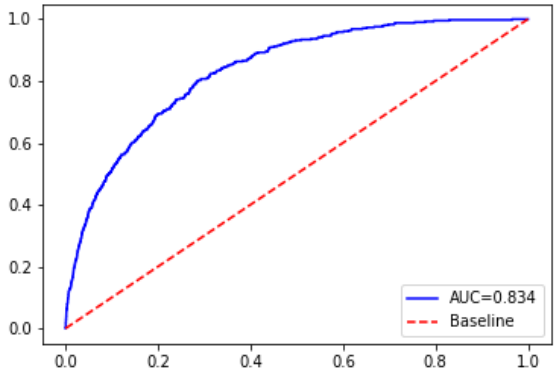


kernel-SVM(AUC = 0.676)

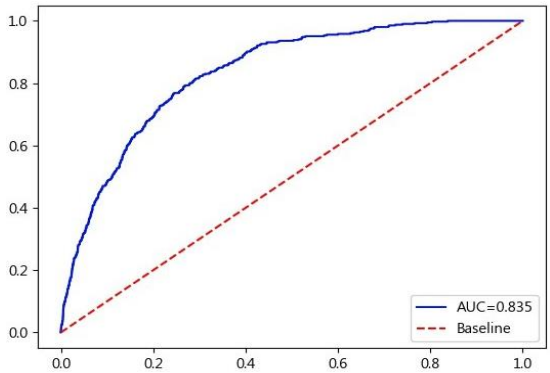


以上圖表為 ML 的各種 Model，明顯 XGBoost 的表現最佳，因此在 ML 中，我們選擇 XGBoost 作為我們之後 FineTune 來跟 DL 模型比較的方法。

MLP(AUC = 0.834)



CNN(AUC = 0.835)



以上圖表為 Deep Learning 的 Model，由上列圖表可以看出 CNN 的表現比 MLP 好一些，因此在 Deep Learning 模型中，我們將兩者都做 FineTune 後再進一步比較差異。

註：在經過資料預處理後，總共的資料維度為 94，因此我們將前後各補上三個 0，使維度成為 100，並將每一個客戶的資料做成一張 10x10 的圖片後，再用 CNN 來處理。

三、模型驗證與 finetune

首先是 XGBoost 的部分，以下大致分三階段做參數 finetune。

第一階段發現設 n_estimators=200、max_depth=3 的 AUC 最佳。

learning_rate	Default=0.1	0.1	0.1	0.1	0.1	0.1
n_estimators	Default=100	200	100	50	200	200
max_depth	Default=6	4	4	4	5	3
min_child_weight	Default=1	1	1	1	1	1
gamma	Default=0	0	0	0	0	0
subsample	Default=1	0.8	0.8	0.8	0.8	0.8
colsample_bytree	Default=1	0.8	0.8	0.8	0.8	0.8
scale_pos_weight	Default=1	1	1	1	1	1
AUC	0.845	0.847	0.846	0.839	0.843	0.848

第二階段最佳 AUC 為 0.849

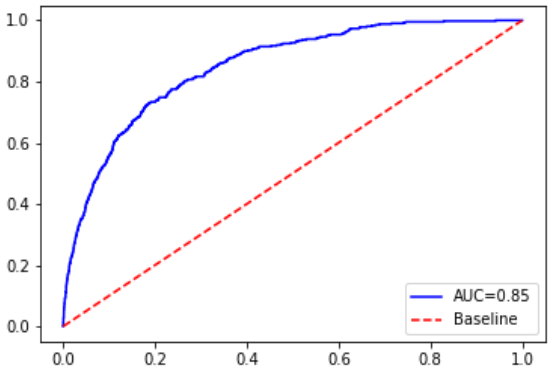
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n_estimators	200	500	200	200	200	200	200
max_depth	2	3	3	3	3	3	3
min_child_weight	1	1	1	3	1	1	1
gamma	0	0	0	0	0	0	0
subsample	0.8	0.8	0.8	0.8	1	0.6	1
colsample_bytree	0.8	0.8	0.8	0.8	0.8	0.8	0.6
scale_pos_weight	1	1	1	1	1	1	1
AUC	0.842	0.845	0.848	0.847	0.847	0.846	0.849

第三階段最佳 AUC 為 0.85

learning_rate	0.1	0.1	0.1
n_estimators	200	200	200
max_depth	3	3	3
min_child_weight	1	1	1
gamma	0	0	0
subsample	1	1	1
colsample_bytree	0.6	0.6	0.6
eval_metric	'auc'	'auc'	'auc'
scale_pos_weight	1	1	1
reg_alpha	0.5	0.7	0.3
AUC	0.85	0.847	0.848

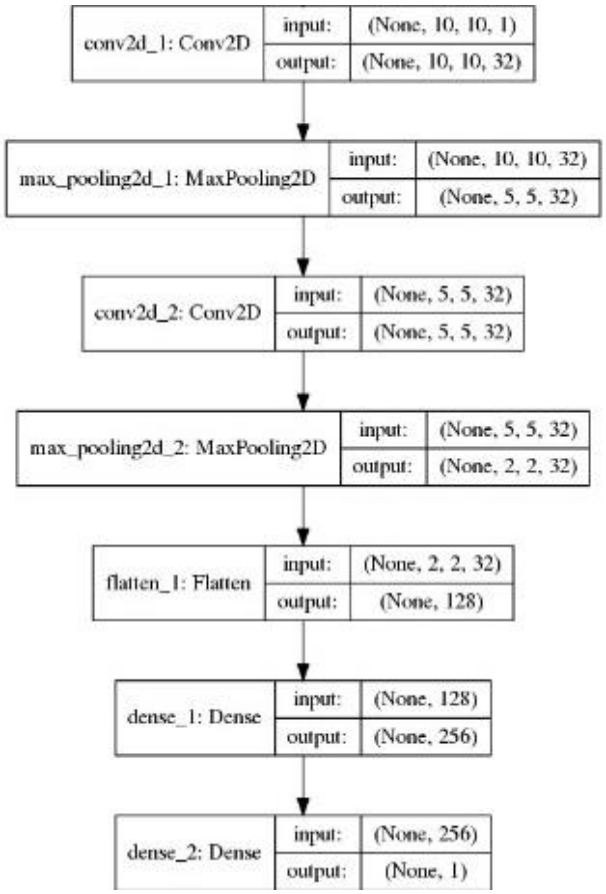
經過多次調整參數的結果，發現 gamma 、scale_pos_weight 這兩個欄位對 AUC 並沒有影響，而調整後的結果 probability=True,

learning_rate=0.1, n_estimators=200, max_depth=3, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.6, reg_alpha=0.5, scale_pos_weight=1 的結果對比全部使用預設值的結果，AUC 會從 0.845 提升到 0.85。



再來是 CNN 的參數測試與 finetune：
下表為實驗各種不同的層數、filter 數和不同的 neuron 數的結果，目的是找出最佳的參數以使得 CNN 能達到最佳的效果。

		全連接層的 neuron 個數		
		128	**256**	**128+128**
Convolution + Max pooling layer 的 filter 數	**32**	0.83	0.828	0.834
	64	0.836	0.832	0.837
	32+32	0.833	0.837	0.833
	32+64	0.831	0.833	0.833
	32+32+64	0.828	0.828	0.831



右圖為最佳的 CNN 模型架構之一。結果顯示一層 convolution 加上兩層 fully connective layer，以及兩層 32 個 filter 的 convolution 加上一層 256 個神經元的 fully connective layer 的表現是最佳的，但是 AUC 的結果還是不如 XGBoost。

四、資料切分與採樣方法測試

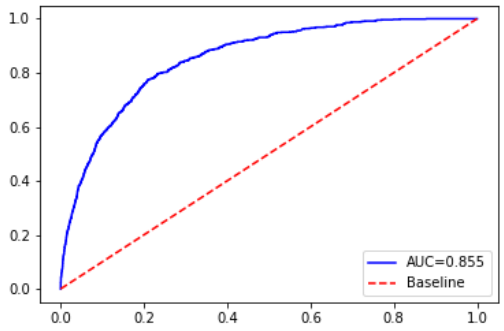
•調整 Train set size

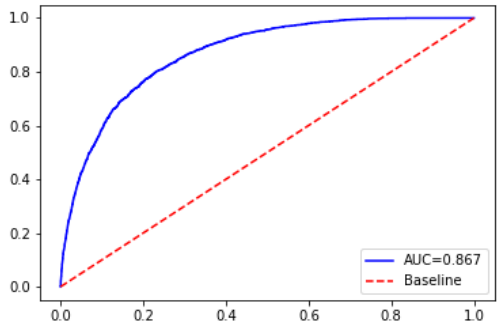
由於本次 training data set 有 10 萬筆客戶資料，我們認為 training set 跟 validation set 的比例對訓練模型會有影響，於是做了資料比例切分的測試，將 validation set 比例從 0.1、0.2、0.3、0.4、0.5、0.6 分別測試，最後發現 validation 比例設成 0.2 最佳(即訓練集有 8

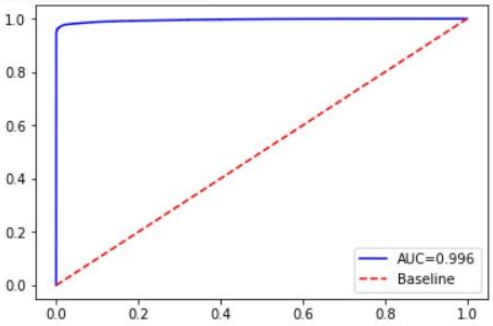
萬筆來餵進 model)，並做 k-fold 後 XGBoost 的 AUC 都有 0.85 以上。若改設成 0.1 或 0.3 其 AUC 大約 0.84 左右，若設成 0.4 以上其 AUC 甚至會掉到 0.826。

•過採樣方法測試 Oversampling

由於 Dataset 中，客戶在未來三個月內是否會購買重疾險保單之比例相差相當懸殊，只有 2%會購買屬典型非平衡數據預測問題，所以我們嘗試了幾個 oversampling 的方式，以下各圖是 XGBoost 的結果。右圖是第一種方法，是將有購買重疾險之客戶資料複製並加入訓練集中。重新訓練模型之後，發現此操作對於 validation set 之 AUC 上升到 0.855，但在測試集的結果上傳後卻下降到 0.8408，推測是 Overfitting 造成的結果。我們仍嘗試降低複製的數量，結果仍然是不佳，於是我們判斷傳統複製型的過採樣方法在此次競賽不可行。





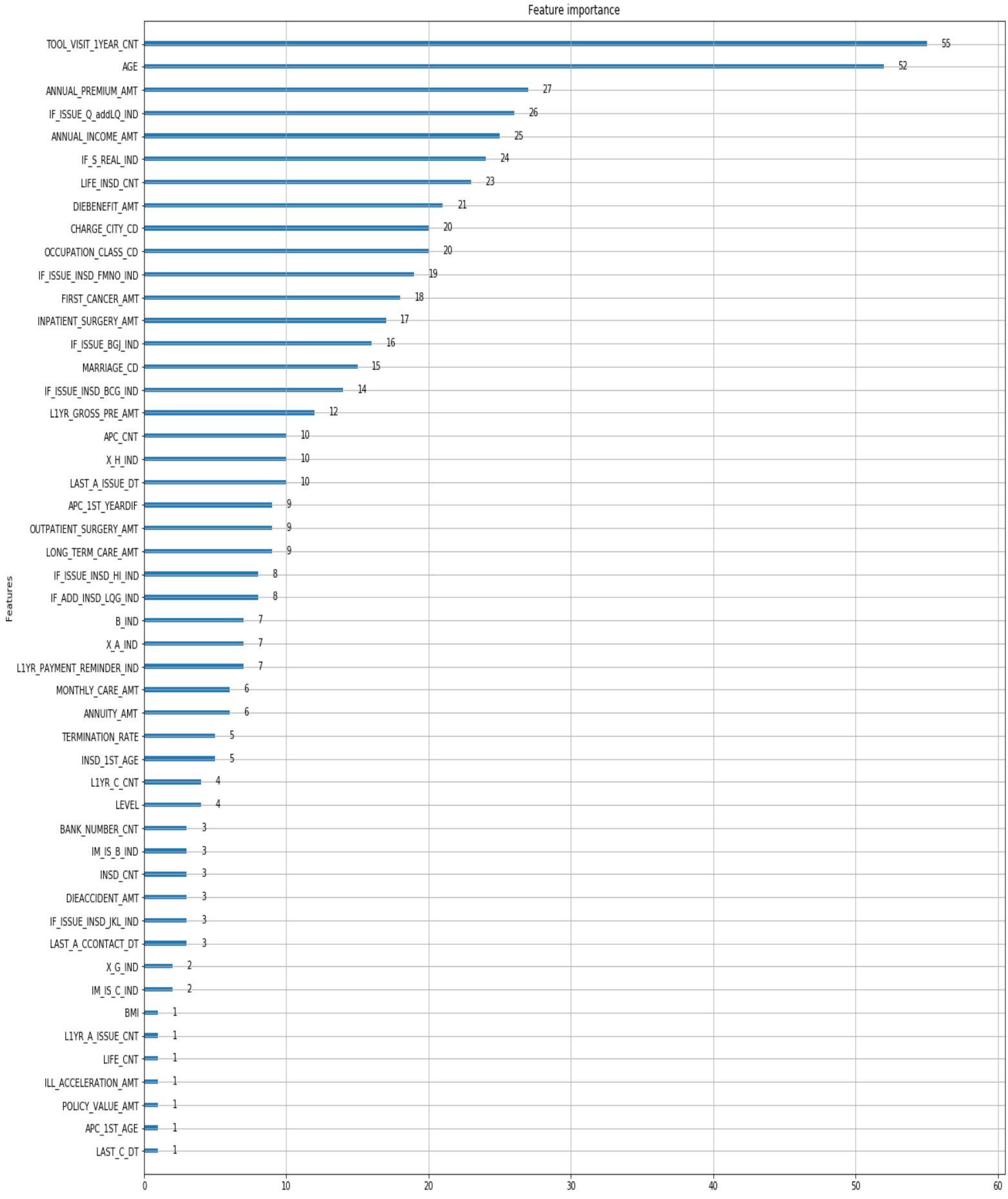


最後我們再嘗試了「隨機樸素過採樣」和「自適應合成過採樣」（ADASYN），其 validation set 結果如上圖。左圖是隨機樸素過採樣的結果，其 AUC=0.867，然而其上傳結果為 0.8405。右圖為 ADASYN 的結果，其 AUC 高達 0.996，然而其上傳結果 0.8027，明顯都 overfitting，所以看來過採樣方法在這次競賽中沒有幫助。(註：CNN 也嘗試過以上過採樣的方法，整體變化程度跟 XGBoost 差不多，而最後也是沒特別正向的成效。)

五、結論

資料預處理中，我們盡可能降低維度，同系列欄位能合併就合併，各欄幾乎都轉成連續數據，並盡可能保有該有的分類數據特質。模型選擇與驗證部分，我們最終認為 XGBoost 的性能最佳。另外，我們也再嘗了幾種過採樣的方法，但在 public board 的結果都不升反降，所以我們認為過採樣在此資料集容易 overfitting。

附件一、XGBoost 的 feature importance



附件二、XGBoost Tree Plot