

Machine Learning

Dataset

Competition

Discussion

2019 國泰大數據競賽— NCNUIM 隊



目錄

1. 資料處理與特徵選取	3
1.1 Data Processing	4
1.2 Feature Selection	5
1.3 Data Cleaning.....	6
2. 模型選擇與驗證成效說明.....	7
2.1 Model Selection.....	7
3. 結論.....	9



1. 資料處理與特徵選取

1.1 Data Processing

初步拿到資料先查看資料的概況，發現有十萬筆資料，130 個變數 1 個依變數，有多個自變數缺失達 50%以上，最嚴重的缺失高達 62.7%，依變數(以下統稱 Y)的比例為 98：2，這個資料為高度不平衡的二元分類資料，挑掉的變數如圖 1 所示：

```
IF_ADD_INSD_F_IND      48152 non-null object
IF_ADD_INSD_L_IND      48152 non-null object
IF_ADD_INSD_Q_IND      48152 non-null object
IF_ADD_INSD_G_IND      48152 non-null object
IF_ADD_INSD_R_IND      48152 non-null object
IF_ADD_INSD_IND        99829 non-null object
L1YR_GROSS_PRE_AMT     100000 non-null float64
CUST_9_SEGMENTS_CD     100000 non-null object
FINANCETOOLS_A         37359 non-null object
FINANCETOOLS_B         37359 non-null object
FINANCETOOLS_C         37359 non-null object
FINANCETOOLS_D         37359 non-null object
FINANCETOOLS_E         37359 non-null object
FINANCETOOLS_F         37359 non-null object
FINANCETOOLS_G         37359 non-null object
Y1                     100000 non-null object
dtypes: float64(30), int64(11), object(90)
memory usage: 100.7+ MB
```

圖 1：缺失高達 50%以上的變數

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100000 entries, 3418 to 1994377
Data columns (total 131 columns):
GENDER      99317 non-null object
AGE         100000 non-null object
CHARGE_CITY_CD  100000 non-null object
CONTACT_CITY_CD  100000 non-null object
```

圖 2：總共十萬筆資料

高度不平衡的資料不能看準確率，不平衡資料準確率的高低不是判斷模型的好方法，例如以公司的角度思考，挑出全部的潛在購買客戶的模型會比高準確率但幾乎沒有預測出潛在購買客戶的模型更有價值。

因此以下以 Y 的召回率(Recall)作為判斷模型好壞的主要指標，挑出大部分購買重疾險商品的潛在客戶。

我們認為原始的資料對業務員更容易判斷客戶是否有可能購買保險商品，分析結果可以教導第一線業務員潛在購買客戶具有什麼樣的特徵，因此我們決定不產生新變數來預測客戶。

1.2 Feature Selection

因為資料含有連續型和文字類別型資料不能直接丟進模型，因此我們對類別型資料進行重新編碼，我們刪掉十二個高達 50% 缺失以上的資料，並刪掉遺漏值保留完整的列資料，總共為 7597 筆。接著由隨機森林(Random Forest)模型挑選代表性變數並印出重要性分數以分數高的變數為優先挑選。

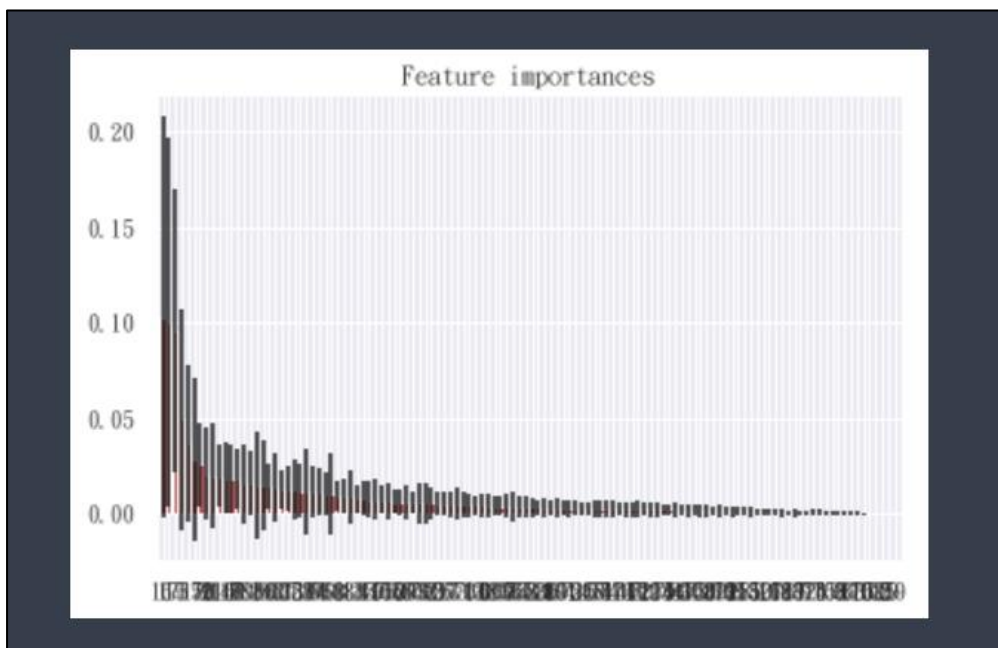


圖 3：變數重要性排行

然後我們選出了排名前 29 的變數，如表 1 所示：

表 1

GENDER	IF_ADD_IND	TOOL_VISIT_1YEAR_CNT
LEVEL	ANNUAL_PREMIUM_AMT	DIEBENEFIT_AMT
EDUCATION_CD	CLC_CUR_NUM	EXPIRATION_AMT
LAST_A_CCONTACT_DT	BANK_NUMBER_CNT	ACCIDENT_HOSPITAL_REC_AMT
L1YR_A_ISSUE_CNT	INSD_LAST_YEAR_DIF_CNT	DISEASES_HOSPITAL_REC_AMT
LAST_B_ISSUE_DT	BMI	OUTPATIENT_SURGERY_AMT
OCCUPATION_CLASS_CD	X_H_IND	PAY_LIMIT_MED_MISC_AMT
APC_CNT	INSD_1ST_AGE	IF_HOUSEHOLD_CLAIM_IND
INSD_CNT	IF_2ND_GEN_IND	IF_ADD_INSD_FAMILY_IND
APC_1ST_AGE	IF_ADD_INSD_L_IND	

接著我們根據這 29 個變數，從原始資料中選取這 29 個沒有缺失的資料，總共有 18716 筆訓練我們的分類模型。

1.3 Data Cleaning

我們經過了很多的資料實驗，我們發現類別型資料用眾數來填補遺漏值，連續型資料用平均數來填補遺漏值，對購買重疾險商品的召回率有明顯的提升。

1	GENDER	EDUCATI	LAST_A_	LIYR_A_	LAST_B_	I	OCCUPAT	AFC_CNT	INSD_CNI	AFC_1ST	INSD_1ST	IF_2ND	G_LEVEL	IF_ADD_I	ANNUAL	CLC_CUR	BANK_NU	INSD_LAS	BMI	X	
2	M	NA	Y			0	N	1	0	0	低	低	N	3	N	0.000192	0	0.125	0.052632	0.125	N
3	M	NA	Y			0	N	1	0	0	低	中	N	5	N	0.009675	0	0.125	0.052632	0.175	N
4	M		1	Y		1	N	1	1	0	低	中	Y	5	N	0.000292	1	0	0	0.225	N
5	M	NA	Y			0	N	1	0	0	低	中	N	5	N	NA	0	0.125	0.210526	0.15	N
6	M		1	N		0	N	1	1	0	低	低	Y	2	N	0.000664	1	0	0.078947	0.175	N
7	M	NA	Y			0	N	1	0	0	低	低	N	3	N	0.000265	0	0.125	0.210526	0.225	N
8	M	NA	N			0	N	1	0	0	低	低	N	3	N	0.002923	0	0.125	0.078947	0.2	N
9	M	NA	Y			0	N	1	0	0	低	低	N	5	N	0.003267	0	0.125	0.052632	0.15	N
10	M		1	N		0	N	1	1	0	低	低	Y	1	N	0.000154	0	0.125	0.131579	0.325	N
11	M	NA	N			0	N	1	1	0	低	低	Y	1	N	0.000201	0	0.125	0.157895	0.3	N
12	M		1	N		0	N	1	1	0	低	中	Y	2	N	0.0011	1	0.125	0.131579	0.2	N
13	M		1	N		0	N	1	2	0	低	低	Y	1	N	NA	0	0.125	0.184211	0.2	N
14	M	NA	Y			0	Y	1	0	0	低	中	N	4	N	0.000256	1	0.25	0.026316	0.25	N
15	M		2	Y		0	N	1	0	0	低	中	N	4	N	0.001604	1	0.125	0.026316	0.25	N
16	M		1	N		0	N	1	1	1	低	低	Y	1	N	NA	0	0	0.078947	0.15	N
17	M	NA	N			0	N	1	0	0	低	中	N	1	N	NA	0	0.125	0.157895	0.2	N
18	M	NA	N			0	N	1	0	0	低	低	N	1	N	0	0	0	0.210526	NA	N
19	M		1	Y		0	N	1	1	0	低	低	Y	5	N	0.019459	1	0	0.078947	0.275	M

原始訓練集 30Xtrain

+

就绪

1	GENDER	EDUCATION	LAST_A_	LIYR_A_	LAST_B_	I	OCCUPATION	AFC_CNT	INSD_CNT	AFC_1ST	INSD_1ST	IF_2ND	G_LEVEL	IF_ADD_I	ANNUAL	CLC_CUR	BANK_NU	INSD_LAS	BMI	X	
2	M	2.169075				0	N	1	0	0	M	低	N	3	N	0.000192	0	0.125	0.052632	0.125	N
3	M	2.169075				0	N	1	0	0	M	中	N	5	N	0.009675	0	0.125	0.052632	0.175	N
4	M	2.169075				1	N	1	1	0	M	中	Y	5	N	0.000292	1	0	0	0.225	N
5	M	2.169075				0	N	1	0	0	M	中	N	5	N	0.001235	0	0.125	0.210526	0.15	N
6	M	2.169075				1	N	1	1	0	M	低	Y	2	N	0.000664	1	0	0.078947	0.175	N
7	M	2.169075				0	N	1	0	0	M	低	N	3	N	0.000265	0	0.125	0.210526	0.225	N
8	M	2.169075				0	N	1	0	0	M	低	N	3	N	0.002923	0	0.125	0.078947	0.2	N
9	M	2.169075				0	N	1	0	0	M	低	N	5	N	0.003267	0	0.125	0.052632	0.15	N
10	M	2.169075				0	N	1	1	0	M	低	Y	1	N	0.000154	0	0.125	0.131579	0.325	N
11	M	2.169075				0	N	1	1	0	M	低	Y	1	N	0.000201	0	0.125	0.157895	0.3	N
12	M	2.169075				0	N	1	1	0	M	中	Y	2	N	0.0011	1	0.125	0.131579	0.2	N
13	M	2.169075				0	N	1	2	0	M	低	Y	1	N	0.001235	0	0.125	0.184211	0.2	N
14	M	2.169075				0	Y	1	0	0	M	中	N	4	N	0.000256	1	0.25	0.026316	0.25	N
15	M	2.169075				0	N	1	0	0	M	中	N	4	N	0.001604	1	0.125	0.026316	0.25	N
16	M	2.169075				0	N	1	1	1	M	低	Y	1	N	0.001235	0	0	0.078947	0.15	N
17	M	2.169075				0	N	1	0	0	M	中	N	1	N	0.001235	0	0.125	0.157895	0.2	N
18	M	2.169075				0	N	1	0	0	M	低	N	1	N	0	0	0	0.210526	0.264241	N
19	M	2.169075				1	Y	1	1	0	M	低	Y	5	N	0.019459	1	0	0.078947	0.275	M

圖 4、圖 5：將原始資料補值

另外由於是做分類預測(Y1 為 Y 與 N)，由於資料要送進預測方法處理，有些分類變數的資料為中文，必須轉為數字的分類資料才可以處理，於是將具有中文的分類變數及其它分類變數一併重新編碼（詳見圖 3），同時將 Y1 變數的 Y 跟 N 改成 1 跟 0，方便我們研究團隊辨識。

1	GENDER	LAST_A_	LAST_B_	I	AFC_1ST	INSD_1ST	IF_2ND_G	IF_ADD_I	X_H_IND	IF_HOUSE	IF_ADD_I	IF_ADD_I	EDUCATI	LIYR_A_	I	OCCUPAT	AFC_CNT	INSD_CNI	LEVEL	ANNUAL_C
2	15	1	1	13	1	17	1	1	9	15	7	1	0	1	1	0	2	0.0011		
3	15	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	5	0.00296		
4	15	5	1	13	1	1	4	1	9	1	1	1	0	1	0	0	4	8.25E-05		
5	15	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	4	0.019695		
6	15	5	1	13	1	1	4	1	1	1	7	1	3	1	0	0	5	0.000307		
7	15	1	1	13	18	17	1	1	9	15	7	1	0	1	1	0	2	7.99E-05		
8	15	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	5	0.002983		
9	15	5	1	13	1	17	1	1	9	1	1	1	0	1	1	0	5	0.000423		
10	15	5	1	13	18	17	1	1	9	15	7	2	0	1	2	0	5	0.005007		
11	15	5	1	13	1	1	4	1	9	1	7	2	0	1	0	0	4	0.000161		
12	15	1	1	13	18	17	1	1	9	1	1	1	0	1	1	0	4	0.000479		
13	15	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	5	0.003052		
14	15	5	1	13	1	17	1	1	1	1	7	2	1	1	1	0	5	0.000966		
15	15	5	1	13	18	17	1	1	9	15	7	2	1	1	2	0	5	5.51E-05		
16	1	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	5	0.001144		
17	1	1	1	13	1	17	4	1	9	1	1	1	0	1	1	0	1	3.06E-05		
18	1	1	1	13	18	17	1	1	9	15	7	1	0	1	1	0	2	0.000957		
19	1	5	1	13	18	17	1	1	9	15	7	1	0	1	1	0	4	0.000459		

圖 6：原始資料補值

至於異常值(Outlier)，雖然對資料來講是 Outlier，但對公司而言，Outlier 有可能是購買商品的潛在客戶，因此我們做保留，希望模型能辨識出，雖然是 Outlier 卻是購買商品的客戶。

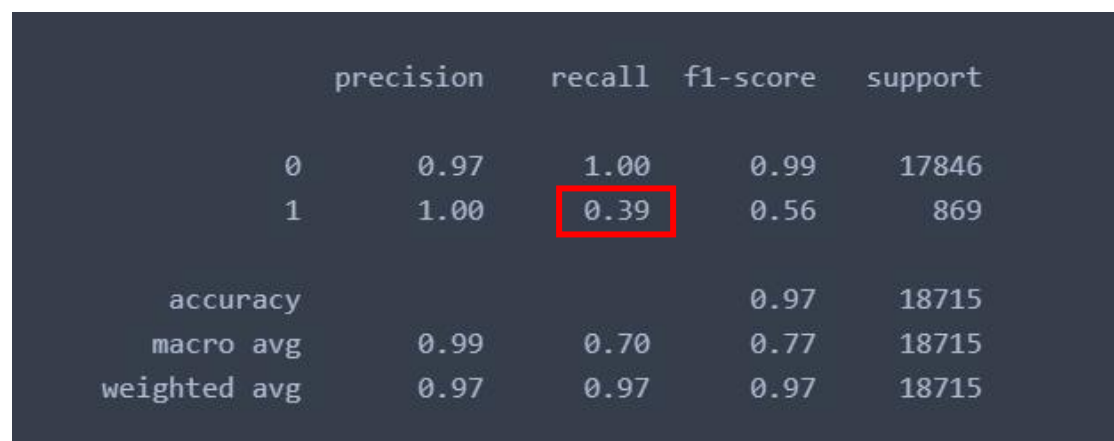
2. 模型選擇與驗證成效說明

2.1 Model Selection

我們希望模型的結果可以讓高層更容易作為決策的參考，而非一堆數據型資料，這些結果可以提供給業務員簡單判斷是否為購買商品的潛在客戶，因此在眾多模型中，我們選擇決策樹(Decision Tree)模型，而決策樹模型又分為許多種，例如：決策樹、隨機森林(Random Forest)，極限樹，這些模型使用不平衡資料建模會失準，因此我們選擇平衡隨機森林 (Balance Random Forest)。

平衡隨機森林 (Balance Random Forest) 訓練過程會使用 Under-Sampling 技術隨機抽取一定相同數量的 0 跟 1 進行訓練，會使用交叉驗證(Cross Validation)，然後我們會把資料切成訓練集跟驗證集且比例是 8：2，因為我們 Train 跟 Test 切成 4：6

圖 7 就是我們隨機森林的結果，在預測會購買的客戶準確率較低，



	precision	recall	f1-score	support
0	0.97	1.00	0.99	17846
1	1.00	0.39	0.56	869
accuracy			0.97	18715
macro avg	0.99	0.70	0.77	18715
weighted avg	0.97	0.97	0.97	18715

圖 7：Random Forest 結果圖

我們可以從圖 7 了解，Accuracy 雖然是 97%，但是在預測 Y=1 的時候 Recall 是不理想的，這代表資料中的 1 都被模型判斷為 0，這樣做是無法找出有可能購買的潛在客戶，所以單純從 Accuracy 是無法替公司增進更多的客戶群。

圖 8 是使用 Balance Random Forest 跑出來的結果，從圖 8 可以發現 Balance Random Forest 在預測的時候，Y=1 的 Recall 指標相較於 Random Forest 明顯提高，能夠挑出大部分的潛在客戶。

	precision	recall	f1-score	support
0	1.00	0.64	0.78	17846
1	0.11	0.93	0.20	869
accuracy			0.65	18715
macro avg	0.55	0.79	0.49	18715
weighted avg	0.95	0.65	0.75	18715

圖 8：Balance Random Forest 結果圖

最後我們使用窮舉方法，最佳化 Balance Random Forest 的分割層數，如圖 9 所示，分割 12 個層數的時候 Y=1 的 Recall 誤差值會達到最小。

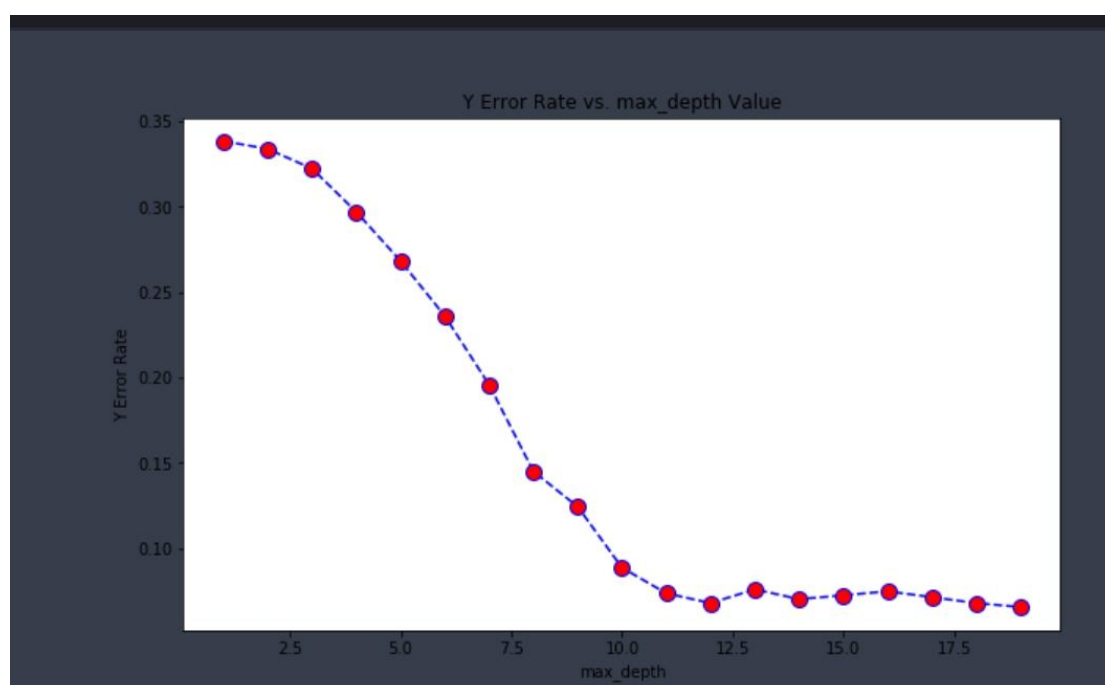


圖 9：窮舉 loss 圖

3. 結論

我們使用 Random Forest & Balance Random Forest 訓練這個資料集，結果證明 Balance Random Forest 演算法雖然只有 75%的 Accuracy，可是可以挑出 93%的潛在客戶，並且其他判斷錯誤的也有跟潛在客戶的特徵相符，業務員如果多加注意，也有可能成為櫃公司客戶。

- 首次擔任被保人年齡(級距)[INSD_1ST_AGE] = 中,低 不符合這個條件的時候購買保險的人較多。
- 首次擔任要保人年齡[級距][APC_1ST_AGE] = 低 不符合這個條件的時候購買保險的人較多
- 婚姻狀況 [MARRIAGE_CD < 1]為符合這個條件的時候 不購買保險的人較多

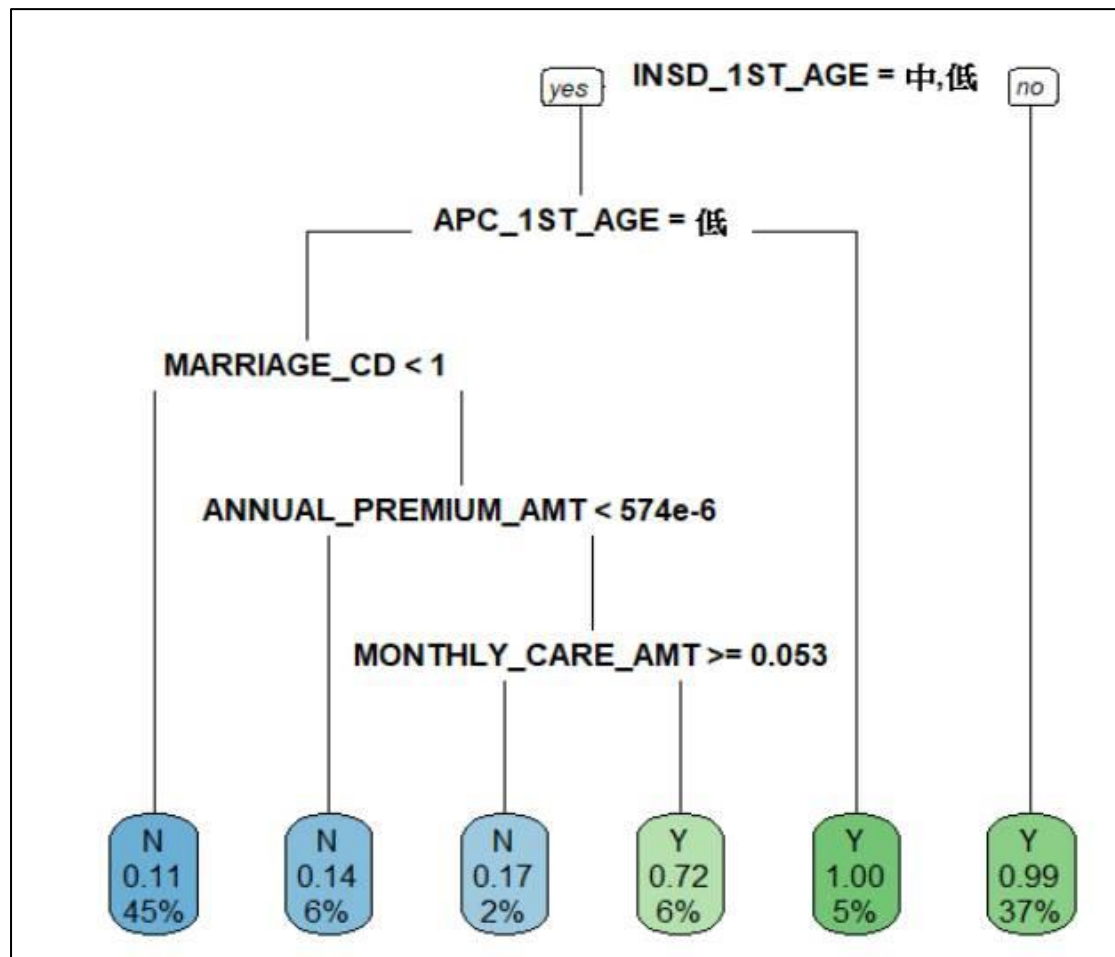


圖 10：Cart 特徵結構圖