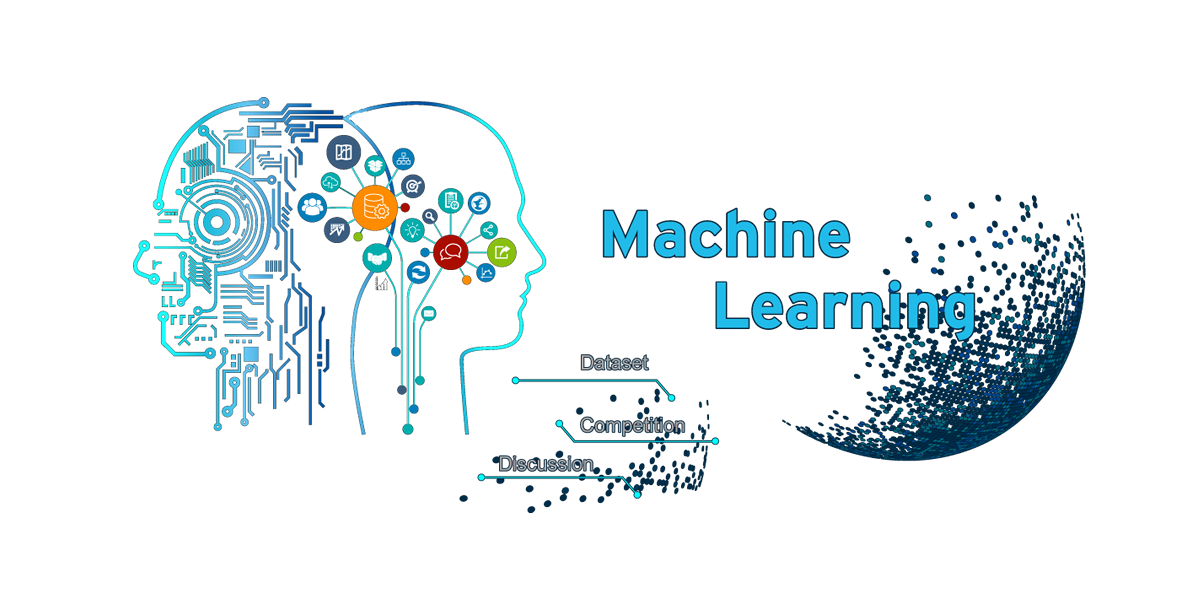
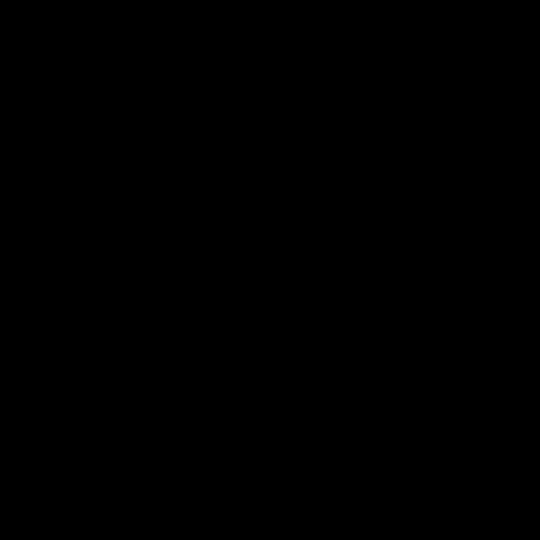
2019國泰大數據競賽－ NCNUIM隊





目錄

1.資料處理與特徵選取 **3**

1.1 Data Processing 4

1.2 Feature Selection5

1.3 Data Cleaning6

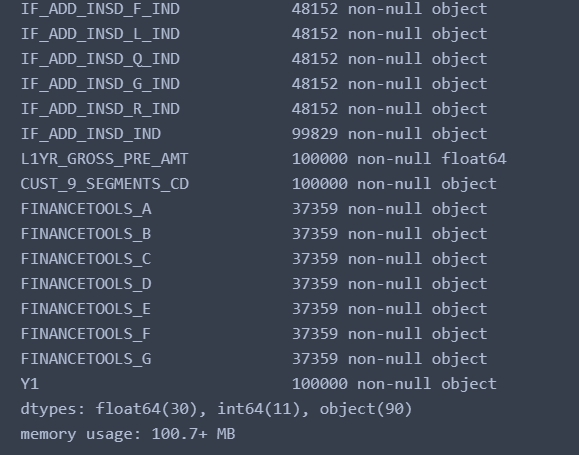
2.模型選擇與驗證成效說明**7**

2.1 Model Selection7

3.結論**9**



1. 資料處理與特徵選取
   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：

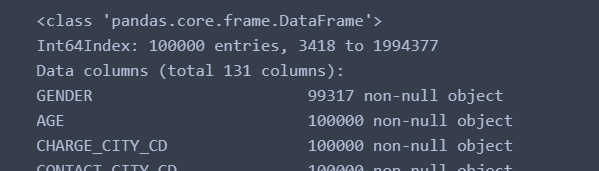
圖1 ： 缺失高達50%以上的變數

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. Feature Selection

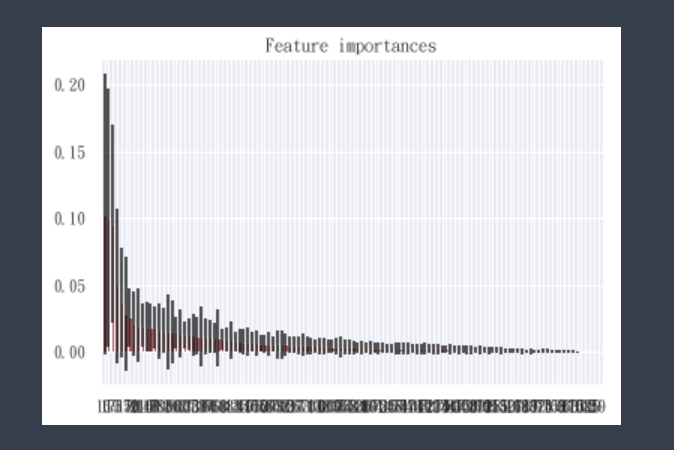
因為資料含有連續型和文字類別型資料不能直接丟進模型，因此我們對類別型資料進行重新編碼，我們刪掉十二個高達50%缺失以上的資料，並刪掉遺漏值保留完整的列資料，總共為7597筆。接著由隨機森林(Random Forest)模型挑選代表性變數並印出重要性分數以分數高的變數為優先挑選。

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：

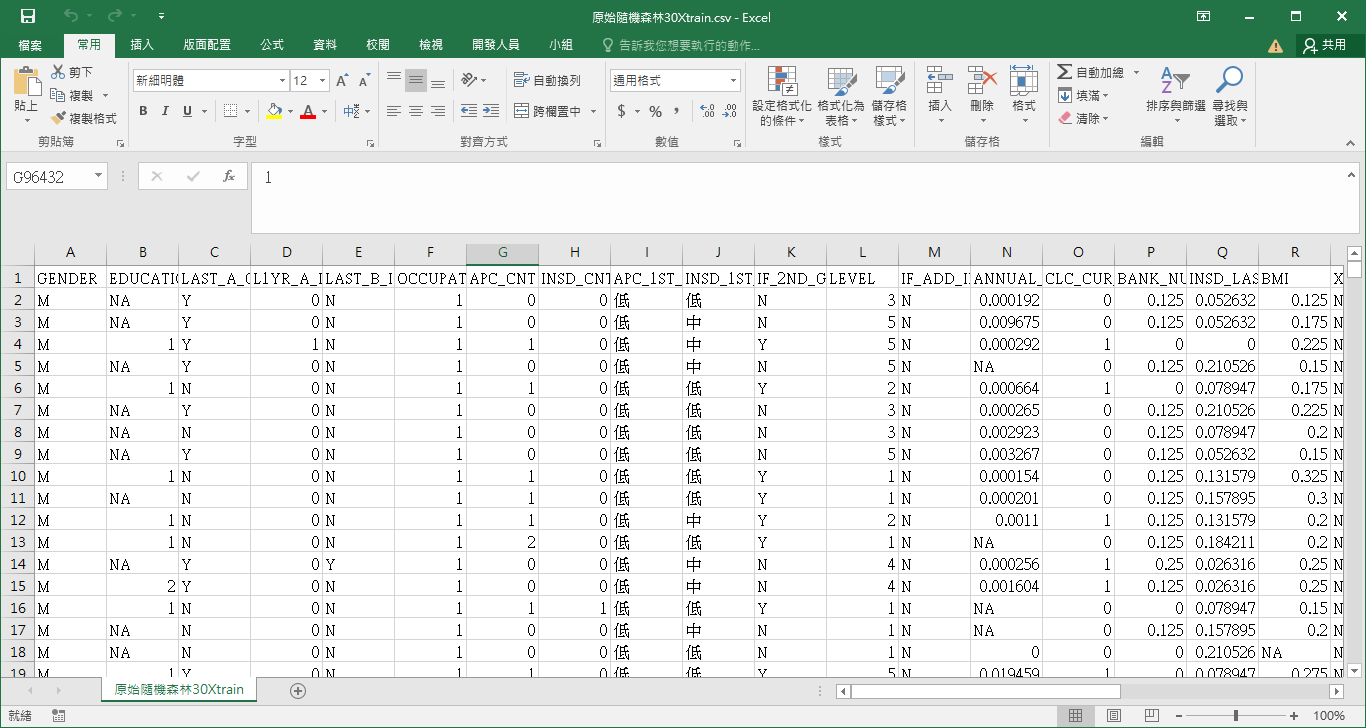
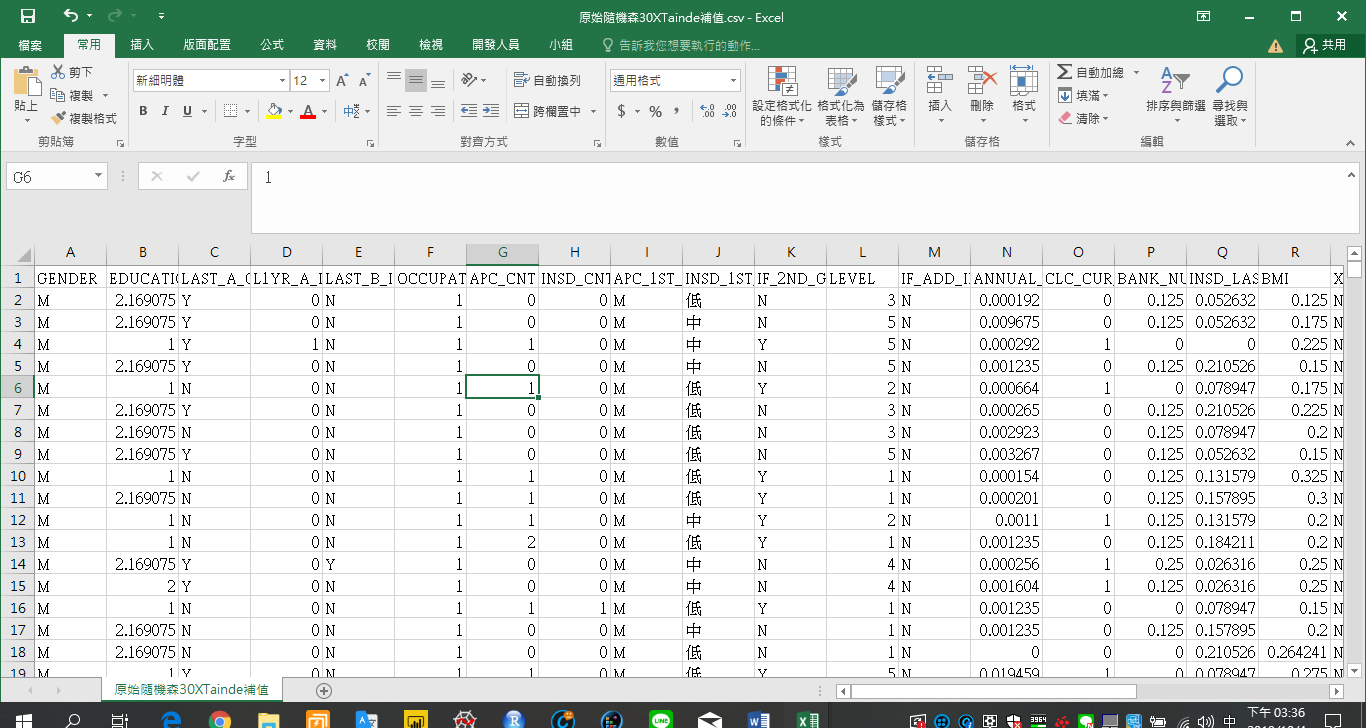
表1

|  |  |  |
| --- | --- | --- |
| 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\_YEARDIF\_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\_F\_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. 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。

圖4、圖5：將原始資料補值

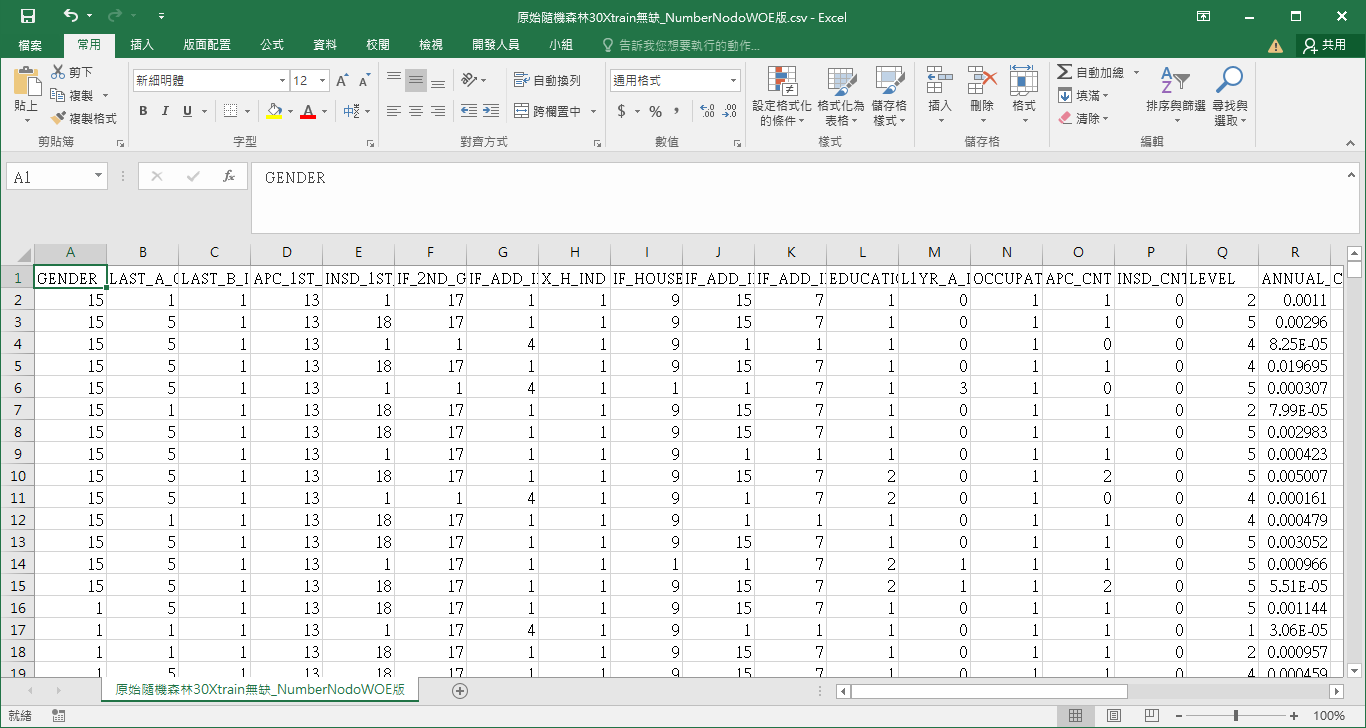
另外由於是做分類預測(Y1為Y與N)，由於資料要送進預測方法處理，有些分類變數的資料為中文，必須轉為數字的分類資料才可以處理，於是將具有中文的分類變數及其它分類變數一併重新編碼（詳見圖3），同時將Y1變數的Y跟N改成1跟0，方便我們研究團隊辨識。

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。

1. 模型選擇與驗證成效說明
   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

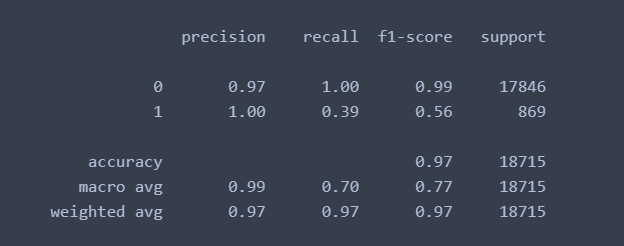
圖7就是我們隨機森林的結果，在預測會購買的客戶準確率較低，

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。

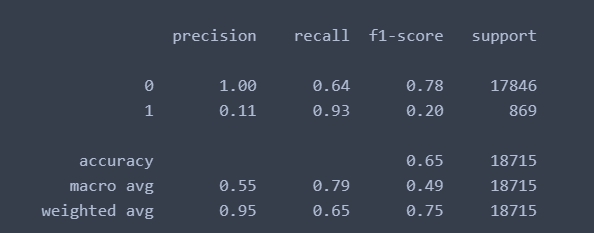
圖8是使用Balance Random Forest跑出來的結果，從圖8可以發現Balance Random Forest 在預測的時候，Y=1的Recall指標相較於Random Forest明顯提高，能夠挑出大部分的潛在客戶。

圖8：Balance Random Forest 結果圖

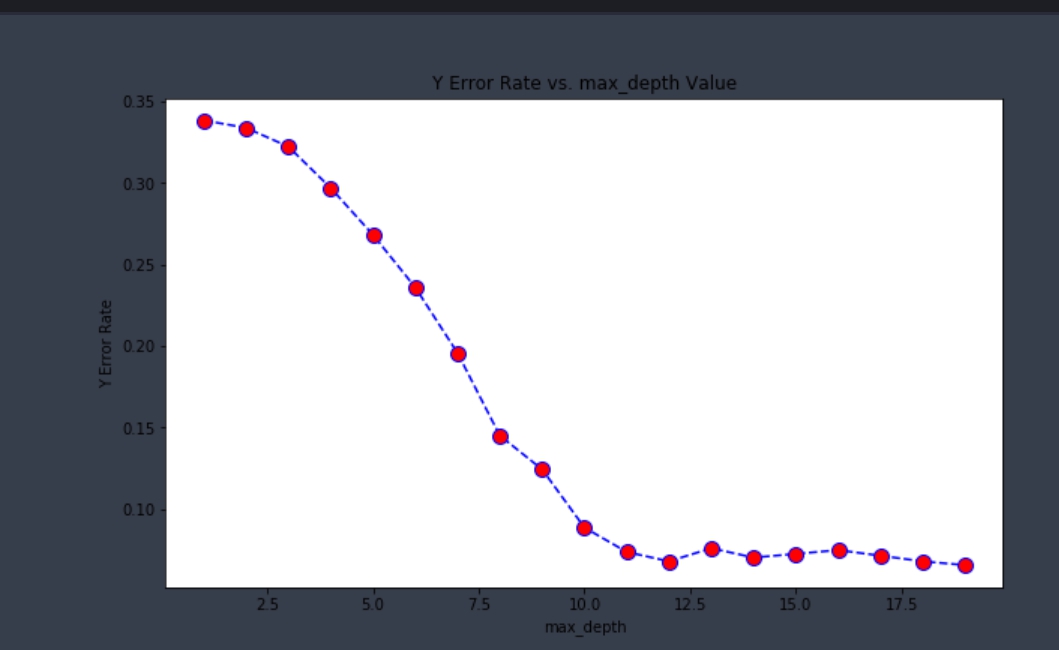
最後我們使用窮舉方法，最佳化Balance Random Forest的分割層數，如圖9所示，分割12個層數的時候Y=1的Recall誤差值會達到最小。

圖9：窮舉loss圖

1. 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。

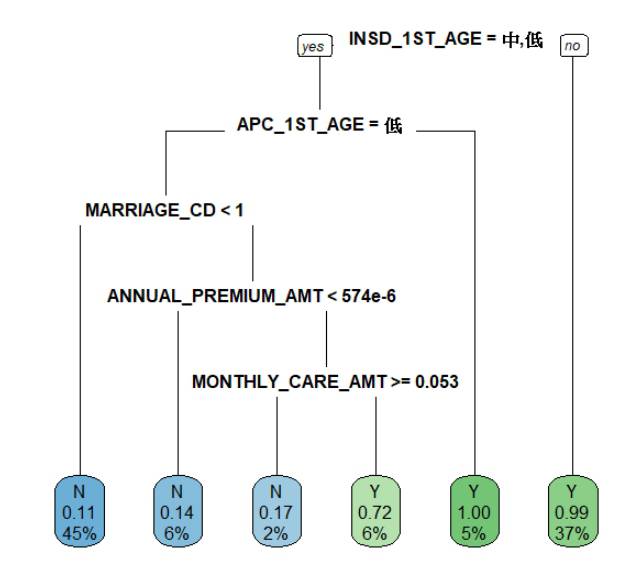
* 首次擔任被保人年齡(級距)[INSD\_1ST\_AGE] = 中,低 不符合這個條件的時候購買保險的人較多。
* 首次擔任要保人年齡[級距][APC\_1ST\_AGE] = 低 不符合這個條件的時候購買保險的人較多
* 婚姻狀況 [MARRIAGE\_CD < 1]為符合這個條件的時候 不購買保險的人較多

圖10：Cart特徵結構圖