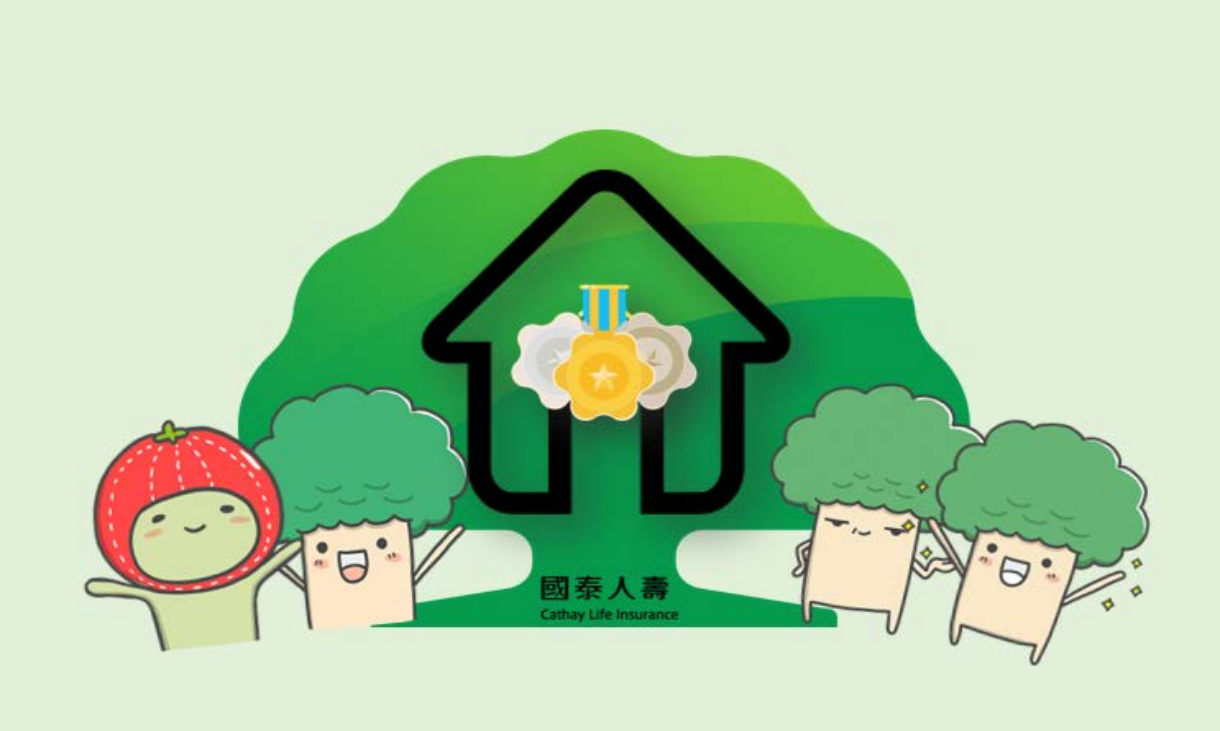
2019

國泰大數據競賽



隊名:冠軍Bang回家

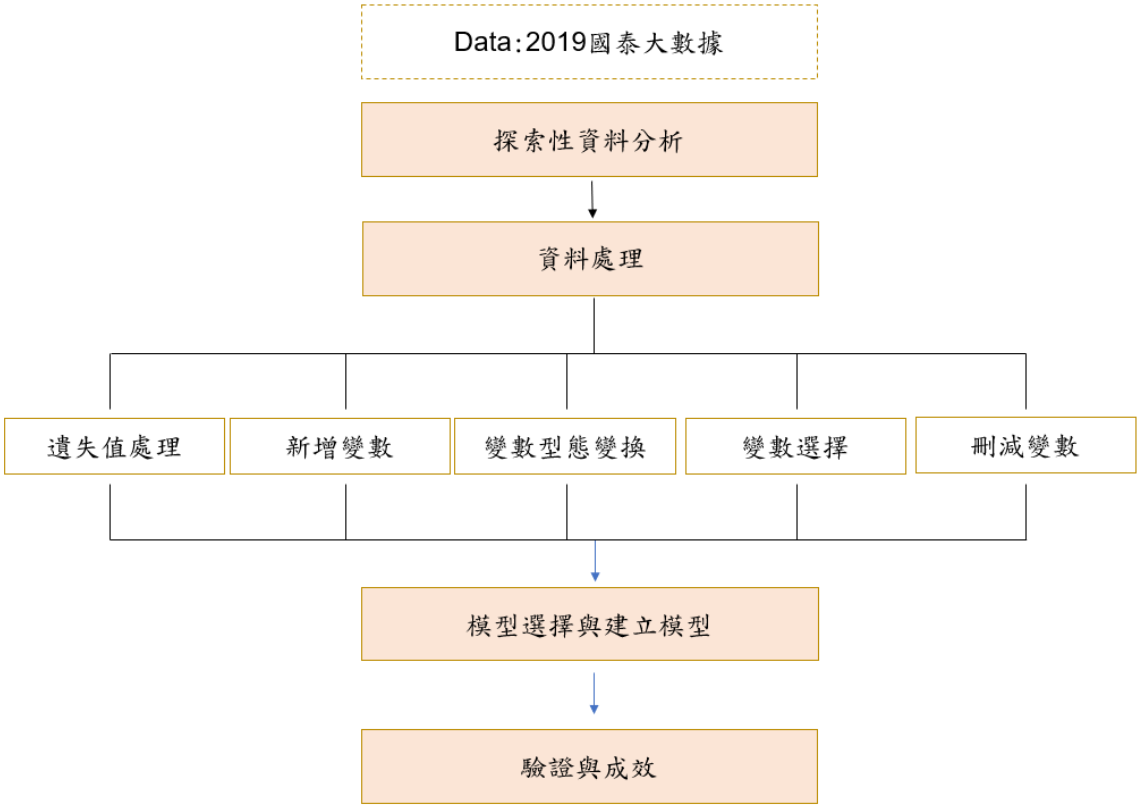
目錄

1. 研究目的

利用既有客戶的購買保單之行為模式等不同特徵去預測未來三個月內是否會購買重疾險商品。此筆資料蒐集既有客戶的相關個人特徵，包含 ： 年齡、職業、當年度保障、收入、年繳化保單等等客戶有關的細項作為預測是否購買重疾險商品的變數。

2. 分析流程圖

分析流程主要有四個階段 ： 探索性資料分析、資料處理、模型選擇、驗證與成效，如下圖(一)。



圖(一) 分析流程圖

3. 資料介紹與敘述

此資料分為Train資料集與Test資料集，資料個數分別為100,000筆與150,000筆。當中，CUS_ID為個人編號，類別型變數共有95個(包含預測變數Y1)，數值型變數共有36個。

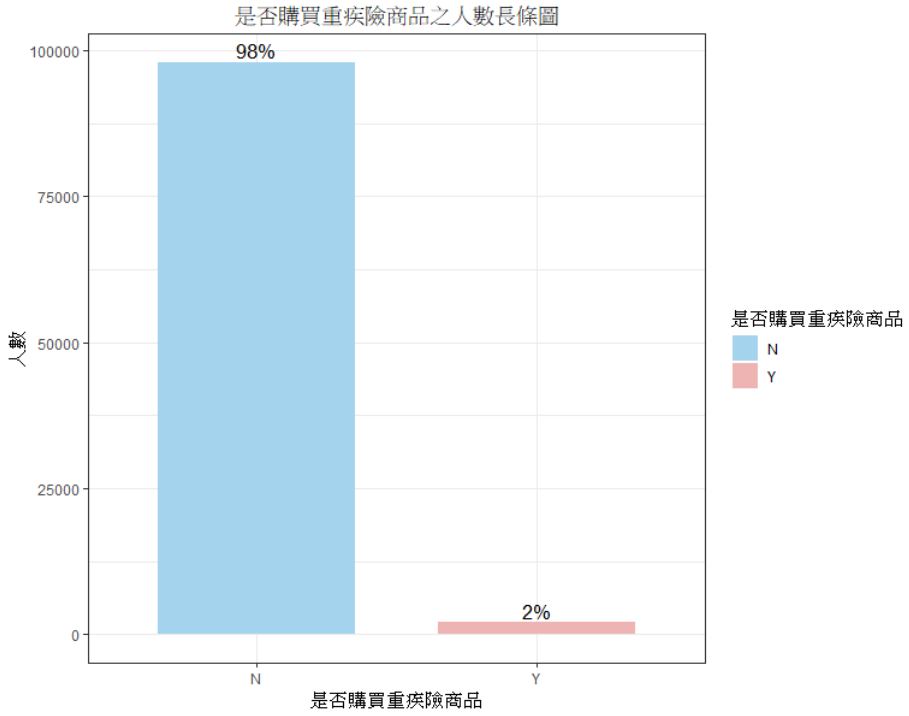
在Train資料集中，預測變數Y1所代表的是「是否有購買重疾險商品」，其類別Y與N的比例懸殊，Y佔2%，N佔了98%(如下表(一))，代表此為不平衡資料，故後續做訓練時的評分標準使用AUC（Area Under Curve)，若是使用準確率進行評斷，則容易陷入因多數類別而導致分類準確率上的誤判。

類別	個數	比例
Y	2,000	2%
N	98,000	98%

表(一) 是否購買重疾險(預測變數)之個數與比例

4. 探索性資料分析

- 預測變數(是否購買重疾險商品)

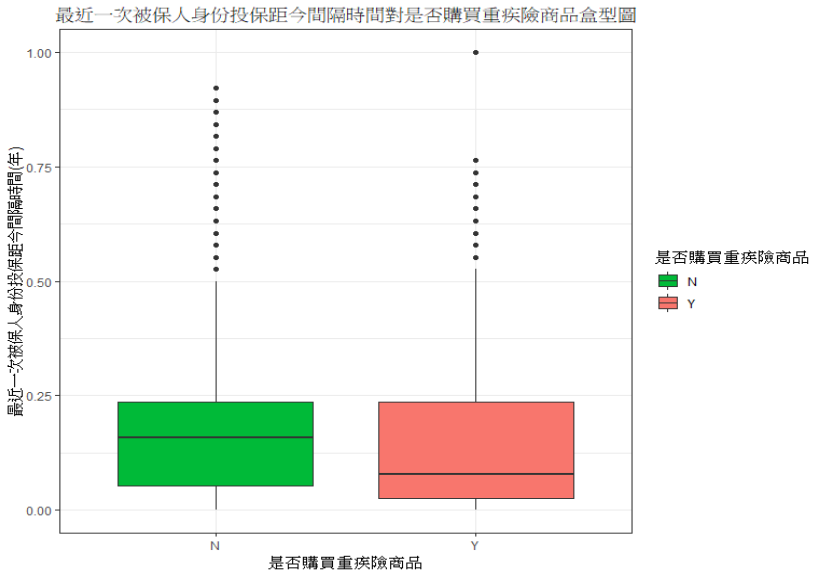


圖(二) 是否購買重疾險商品之人數長條圖

- 解釋變數

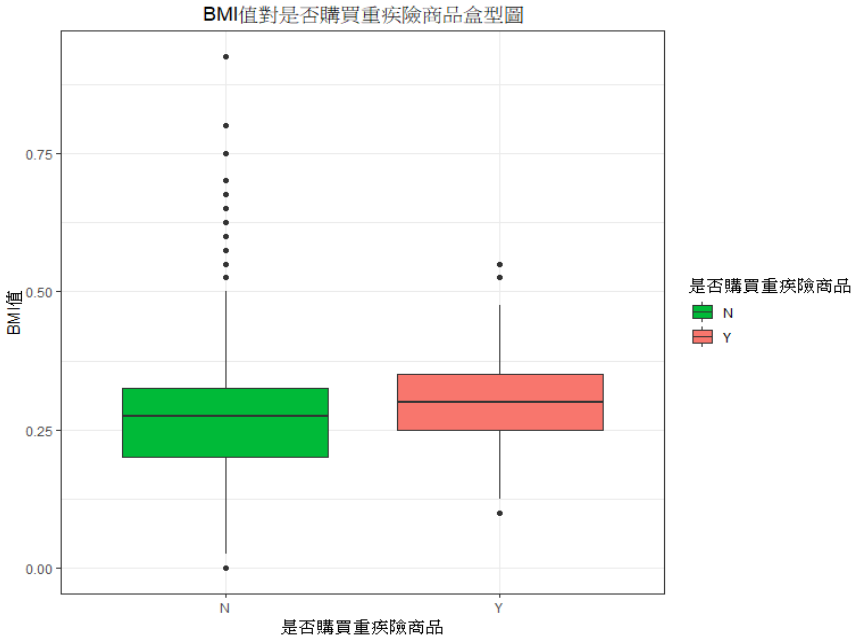
由於變數共有132個(包含預測變數Y1)，無法一一列舉各個變數的樣態並深入分析，也無法隨意主觀判定哪個變數較為重要。因此，我們初步欲利用Train資料集的Y1(是否有購買重疾險商品)當作基礎，觀察有無購買重疾險的這兩類客戶在各個因素中的傾向與嗜好等，以下列舉幾項變數與預測變數Y1之盒型圖與直方圖：

以圖(三)為例，我們可以明顯看出會購買重疾險的客戶，最近一次被保人身份投保距今間隔時間的第25百分位數及中位數較不會購買重疾險的客戶低，可以推斷會購買重疾險的客戶會更頻繁的投保。



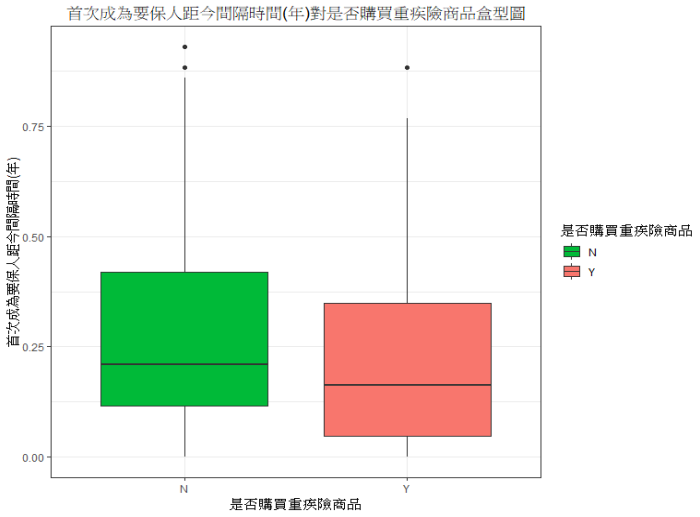
圖(三) 最近一次被保人身份投保距今間隔時間 對
購買重疾險商品客戶 盒型圖

從(圖四)我們可看出會購買重疾險的客戶在BMI的表現:不論是第25、中位數及第75百分位數BMI皆高過不會購買重疾險的客戶，在現有的醫學常識我們知道BMI與健康息息相關，因此BMI會是幫助預測客戶是否購買重疾險商品的重要資訊。



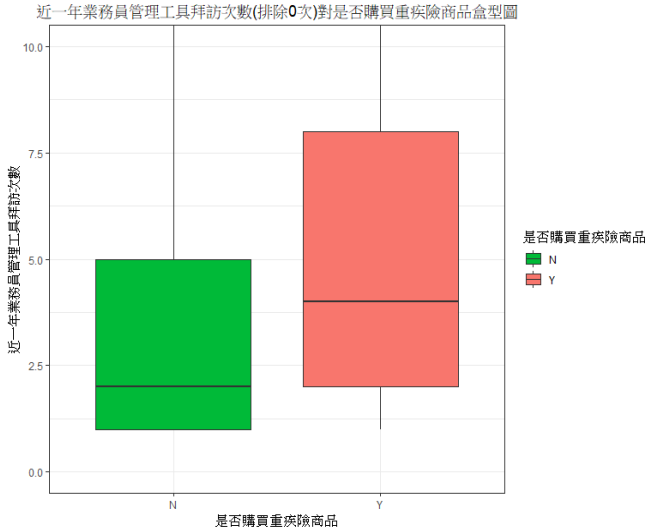
圖(四) BMI 對 是否購買重疾險商品 盒型圖

從圖(五)可以看出會購買重疾險的客戶在首次成為要保人距今間隔時間的長短不論是第25、中位數及第75百分位數的間隔時間皆較不會購買重疾險的客戶短，表示會購買重疾險的那群客戶的保險意識較高，會為自己購買更多保障。



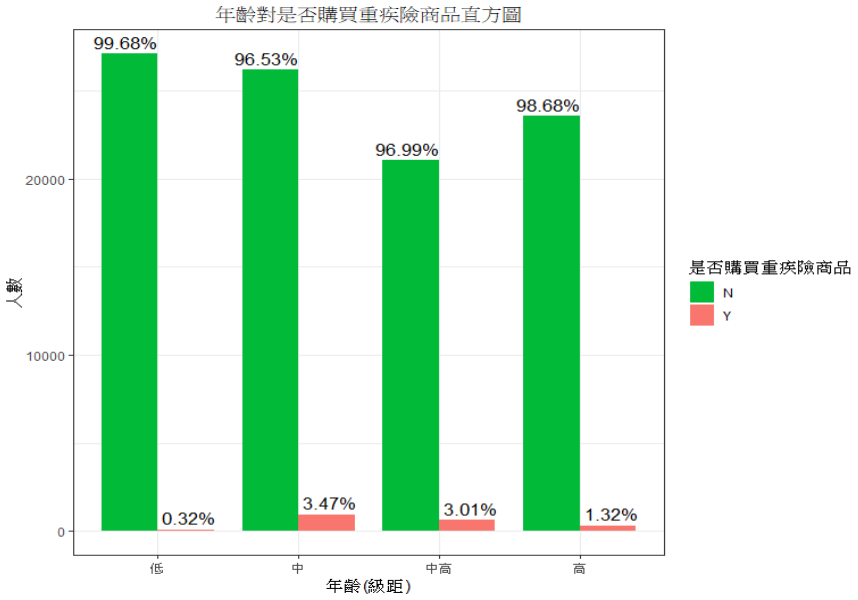
圖(五) 首次成為要保人距今間隔時間(年) 對
是否購買重疾險商品 盒型圖

從圖(六)可以觀察到這個變數包含很多的0，但看到不論是第25、75百分位數及中位數都可以發現會購買重疾險商品的顧客其業務員管理工具的拜訪次數都高於不會買重疾險商品的客戶。可能原因為客戶因為健康疑慮想購買保險所以業務員管理工具拜訪次數比較多，抑或是業務員管理工具拜訪次數較多，所以客戶的購買意願提升。



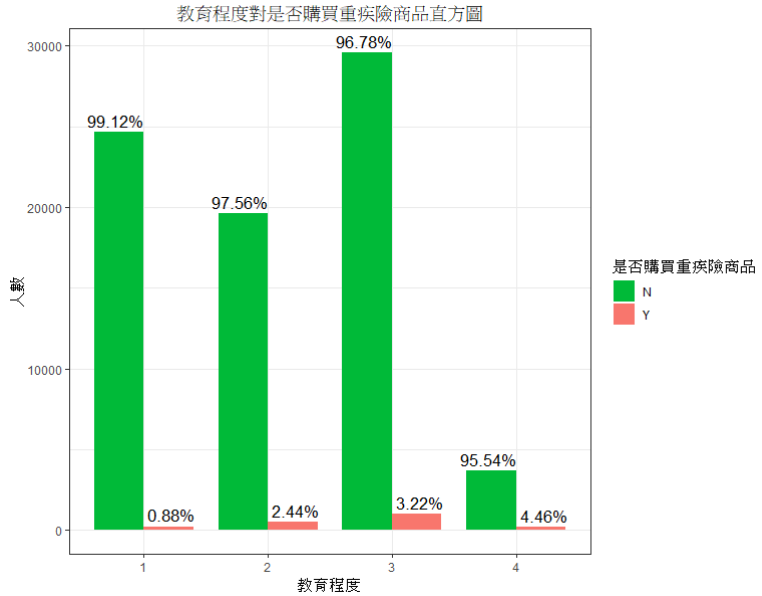
圖(六) 近一年業務員管理工具拜訪次數 對
是否購買重疾險商品 盒型圖

從圖(七)可以看出，因為中年齡層及中高年齡層多為家庭的經濟支柱，故會想為自己購買重疾險商品，減少其得到重疾後的經濟風險。



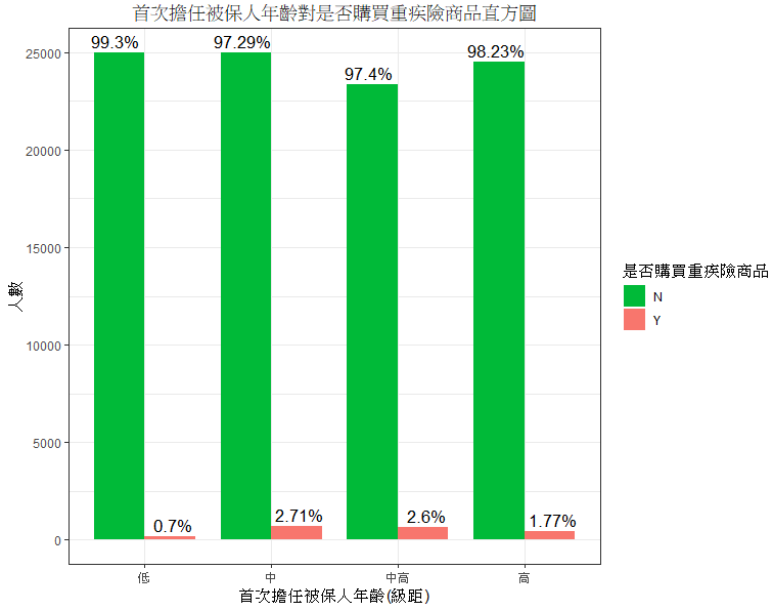
圖(七) 年齡 對 是否購買重疾險商品 直方圖

從圖(八)可以看出隨著教育程度的提高，購買重疾險的比例隨之提升。可以得知其受到教育越多，購買保險的意識也會提升，增加購買重疾險商品的意願。



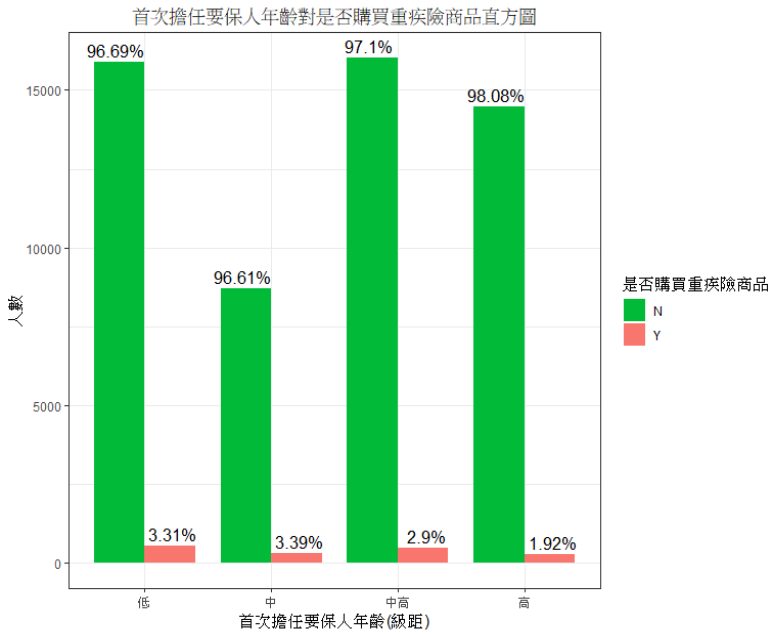
圖(八) 教育程度 對 是否購買重疾險商品 直方圖

從圖(九)可看出首次擔任被保人在年齡等級中及中高的購買重疾險比例較高，造成此現象原因可能與圖(七)類似，由於成為家庭經濟支柱，故責任較大，需要避免得到重大疾病造成的經濟風險。



圖(九) 首次擔任被保人年齡 對 是否購買重疾險商品 直方圖

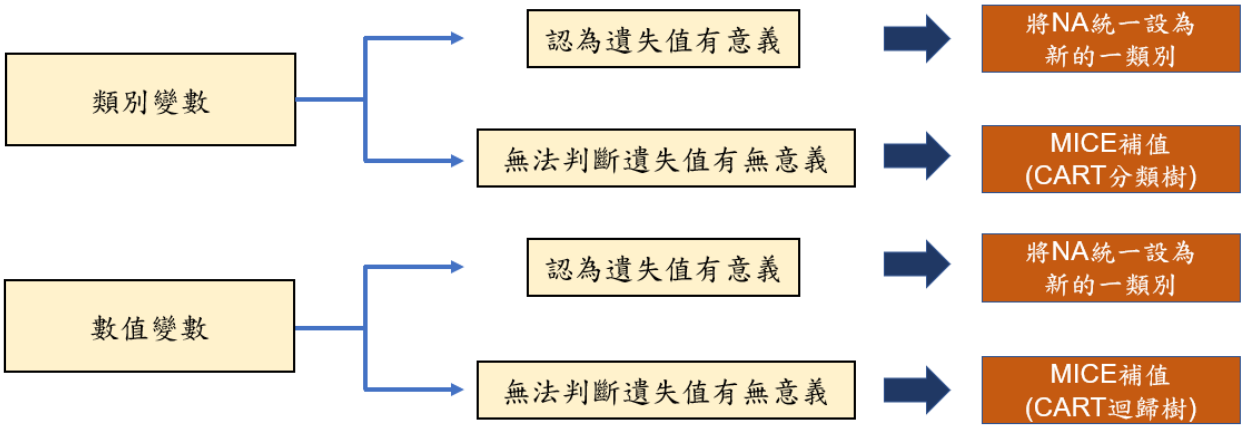
從圖(十)可以看到年齡較低就擔任要保人對於會購買重疾險商品有較高的比例。可以得到因為較早擔任要保人，表示可能很早就接觸有關於保險的知識，故保險的意識較高，購買重疾險商品的比例也提高。



圖(十) 首次擔任要保人年齡 對 是否購買重疾險商品 直方圖

5. 資料處理

- 遺失值補值



在討論過每個變數之後，我們將遺失值分為以下兩種情況。

❖ 第一種情況 ： 在經過討論後，認為遺失值是具有其意義，於是將遺失值設為一個新的類別，名為：NA。

我們觀察到APC_1ST_AGE、APC_1ST_YEARDIF、REBUY_TIMES_CNT、RFM_M_LEVEL、TERMINATION_RATE這五個變數遺失值皆一致，即若一客戶其中一個變數為NA，則另外四個變數也必為NA，而變數本身的涵義皆與要保人身分有關，因此我們判斷在這五個變數出現NA是因為該客戶從未以要保人、皆是以被保人身分購買保單，後續我們以此假設為正確的前提下，對其它資料的遺失值加以詮釋，而模型的預測成績也給了我們良好的反饋，讓我們對此假設的正確性有信心，進而去定義其他變數的遺失值:RFM_R、LEVEL，兩變數在涵義上與要保人身分有關，但遺失值數量有些許差異，因此我們判斷重疊的部分為上述所說的情況，而非重疊的部分為無法歸納原因的遺失值，即是第二種情形(後面會補充會用何種方法填補遺失值)。

	Train遺失值個數	Test遺失值個數
APC_1ST_AGE	43282	64214
APC_1ST_YEARDIF	43282	64214

REBUY_TIMES_CNT	43282	64214
RFM_M_LEVEL	43282	64214
TERMINATION_RATE	43282	64214
RFM_R	43294	64227
LEVEL	43305	64216

而INSD_1ST_AGE、INSD_LAST_YEARDIF_CNT、IF_ADD_INSD_IND則是與被保人身分有關，而同樣地這三個變數不管在Train資料集以及Test資料集出現遺失值也一致，並且變數涵義皆與被保人身分有關，因此我們判斷這三個變數遺失值的出現是因為該客戶從未以被保人、皆是以要保人身分購買保單。

	Train遺失值個數	Test遺失值個數
INSD_1ST_AGE	171	14
INSD_LAST_YEARDIF_CNT	171	14
IF_ADD_INSD_IND	171	14

再來為是否壽險保單被保有效類別有關的變數，即為IF_ISSUE_INSD_A_IND至 IF_ISSUE_INSD_Q_IND共17個與保單被保主約有關的變數以及IF_ADD_INSD_F_IND 至 IF_ADD_INSD_R_IND共5個與保單被保附約有關的變數，這合計22個變數之間有相當有趣的關係，主約附約各自的遺失值出現是一致的，即若一主約類別變數為NA，其他16個主約類別變數皆為NA，附約亦是如此，非重疊部分的客戶我們解讀為當下保單只有被保主約或附約其中一種，因此在另一種被保類別會出現遺失值；而重疊部分的客戶則是目前沒有被保任何保單。

	Train	Test

被保主約類別遺失值個數	20083	29911
被保附約類別遺失值個數	51848	77499
被保主約&附約遺失值個數	19619	29232

再來是與當年度保障相關的變數DIEBENEFIT_AMT、DIEACCIDENT_AMT至 MONTHLY_CARE_AMT共15個變數，這部分的遺失值我們解讀為當年度並沒有該類別的保障(因為數據經神秘轉換，資料顯示為0不見得是保障金額真的為0)。

ANNUAL_PREMIUM_AMT與未以要保人身分投保的客戶有高度的重疊，因此我們判斷年繳化保費的遺失值可解讀為當年度為非要保人，沒有需要繳交的保費。

最後是EDUCATION_CD(教育程度/學歷)、MARRIAGE_CD(婚姻狀況)、ANNUAL_INCOME_AMT(年收入)三個較偏向個人隱私的變數，而客戶之所以不願意填此資料可能有其考量(水準不如平均，或有其他難言之隱)，因此我們認為這兩個變數的遺失值是能夠提供資訊的，因此將遺失值令為新類別。

❖ 第二種情況 ： 在經過討論後，無法判斷遺失值是否具有其意義，於是使用MICE進行補值，利用分類或迴歸分析幫助我們判斷其應該為何值。

● 補值方法介紹（MICE)：

MICE 的基本想法為對一個具有遺失值的變量，利用其他變量的數據對於這個變量進行建模預測分析，將其預測分析的最終結果將其帶入遺失值的欄位進行補值，對於數值型變數，MICE會利用CART迴歸樹進行分析預測，若是類別型變數，則MICE會利用CART分類樹進行預測分析。

● 變數型態變換

為了反應某些變數欄位其順序型的資訊，因此分別將其欄位內容改為1、2、3、4，及1、2、3。

變數	內容轉換

AGE、INSD_1ST_AGE、RFM_R、REBUY_TIMES_CNT、APC_1ST_AGE	低、中、中高、高 → (1,2,3,4)
LIFE_CNT	低、中、高→(1,2,3)

我們將兩個變數的型態做了變換，包括EDUCATION_CD、MARRIAGE_CD，我們觀察到這兩個變數皆屬於類別變數的範疇，但原檔案將其設置為浮點數(float)，也就是包含小數點的數值型型態，因此透過程式碼，將其皆轉換為類別變數(object)，其餘變數型態皆不變。

● 變數選擇

在變數選擇上，由於我們先對遺失值都進行了填補或將其設成一類別，因此我們想先嘗試使用全部的變數去執行我們所考慮的所有模型，其結果也達到了相對不錯的評分，而在後續，我們進行了重要變數選取，將其中幾個重要性係數較高的變數選取出來後，並且刪除些許變數與新增兩項變數後，再次進行了模型訓練。

● 變數刪除

經討論後，將歸納不出其意義之變數，且其NA稍多(10%)之變數，將之刪除，刪除的變數有以下：

A_IND、B_IND、C_IND、L1YR_C_CNT、FINANCETOOLS_A、FINANCETOOLS_B、FINANCETOOLS_C、FINANCETOOLS_D、FINANCETOOLS_E、FINANCETOOLS_F、FINANCETOOLS_G

● 變數新增

1、在判斷NA意義時，認為是否成為**被保人**與**要保人**之身份確認是對於分類較為重要的資訊，為此新增兩變數為NEVER_APC、NEVER_INSD，以提供模型此資訊。

2、我們發現CHARGE_CITY_CD(收費地址)與CONTACT_CITY_CD(聯絡地址)不一定相同，可能導致聯絡方與收費方牽扯到不同人的情形，導致接受到的資訊因此不同，又或是聯絡方為被保人、收費方為要保人的情形，故新增一變數為CITY_SAME，判斷其地址是否為一致，以提供模型此資訊。

		是否購買重疾險商品	
		Y	N
是否成為要保人	Y	387 (19.35%)	42895 (43.77%)
	N	1613 (80.65%)	55105 (56.23%)

		是否購買重疾險商品	
		Y	N
是否成為被保人	Y	64 (3.2%)	107 (0.11%)
	N	1936 (96.8%)	97893 (99.89%)

		是否購買重疾險商品	
		Y	N
聯絡地址是否與收費地址相同	Y	1704 (85.2%)	64349 (65.66%)
	N	296 (14.8%)	33651 (34.34%)

6. 模型選擇與驗證成效說明

- 模型選擇

  我們考慮使用CatBoost模型去分析資料，而這種Boosting的方法皆屬於Gradient Boosting Decision Tree（GBDT） 的方法延伸，而GBDT顧名思義即為梯度提升決策樹，其做法為在每次的疊代中，都是去建立新的殘差預測，一次次的加入到上一次更新的預測值上，使其與原先的標籤值之間的殘差越來越小，以達到預測準確提高的目的。

  至於為何會使用Boosting方法是因為其也會在某些程度上的解決此次資料不平衡的問題，因Boosting在每次的疊代中，都會將此次預測不準的資料點進行權重加權，使其更容易的被分類器選中，就不會有資料不平衡時，一般將少數類別當作離群值忽略的情況發生了。

- 方法介紹與選擇理由

  ✓ Catboost

  CatBoost 為近年來較為新穎的方法，其在類別變數的處理上，做了非常大的優化，CatBoost 在進行類別變數的處理上，CatBoost 除了使用One hot encoding 處理以外，在針對設定上超過若干數量類別的類別變數時，CatBoost 會使用另一種處理方法，將所有輸入的資料隨機排列後，進行有序增強，將類別特徵值轉為數值，而CatBoost 在數值型變數的處理上，就與其他決策樹相同，使用信息增益選擇最佳拆分，並且CatBoost 在數據的處理上，能夠同時處理數值型資料與類別型資料，因此對於如本筆遺失值較多的

資料，無法界定遺失值如何判別時，可先將遺失值設為一類，CatBoost 模型依舊能夠處理，這也是為何會選擇CatBoost 的一個重要的原因。

- 驗證成效

在最終提交的結果中，CatBoost的回傳結果幾乎都是在84.40%以上，其中最高的一次是落於85.138%。

附錄

- 解釋變數的遺失值處理

認NA為新類別的變數					
EDUCATION_CD	MARRIAGE_CD	APC_1ST_AGE	INSD_1ST_AGE	APC_1ST_YEARDIF	IF_ADD_INSD_IND
RFM_R	REBUY_TIMES_CNT	LEVEL	RFM_M_LEVEL	ANNUAL_PREMIUM_AMT	EXPIRATION_AMT
ANNUAL_INCOME_AMT	INSD_LAST_YEARDIF_CNT	TERMINATION_RATE	DIEBENEFIT_AMT	DIEACCIDENT_AMT	ANNUITY_AMT
POLICY_VALUE_AMT	FIRST_CANCER_AMT	INPATIENT_SURGERY_AMT	ACCIDENT_HOSPITAL_REC_AMT	DISEASES_HOSPITAL_REC_AMT	PAY_LIMIT_MED_MISC_AMT
OUTPATIENT_SURGERY_AMT	MONTHLY_CARE_AMT	ILL_ADDITIONAL_AMT	LONG_TERM_CARE_AMT	ILL_ACCELERATION_AMT	IF_ISSUE_INSD_A_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_B_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_ADD_INSD_R_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	IF_ADD_INSD_G_IND

IF_ADD_INSD_F_IND	IF_ADD_INSD_L_IND	IF_ADD_INSD_Q_IND			

MICE補值的變數					
GENDER	OCCUPATION_CLASS_CD	BMI	X_A_IND	X_B_IND	X_C_IND
X_D_IND	X_E_IND	X_F_IND	X_G_IND	X_H_IND	

- 參考資料

1、特徵工程到底是什麼？
https://tinyurl.com/y65hup6c
2、Feature Engineering 特徵工程中常見的方法
https://tinyurl.com/y2yxww2w
3、CatBoost - open-source gradient boosting library
https://catboost.ai
4、XGBoost Documentation
https://xgboost.readthedocs.io/en/latest/
5、ggplot2 Quick Reference: colour (and fill)
https://tinyurl.com/y6fdl9x5
6、Changing the order of levels of a factor
https://tinyurl.com/owsjkof