技安家庭

分析說明書

Methods Analysis Specification

2019 國泰大數據競賽

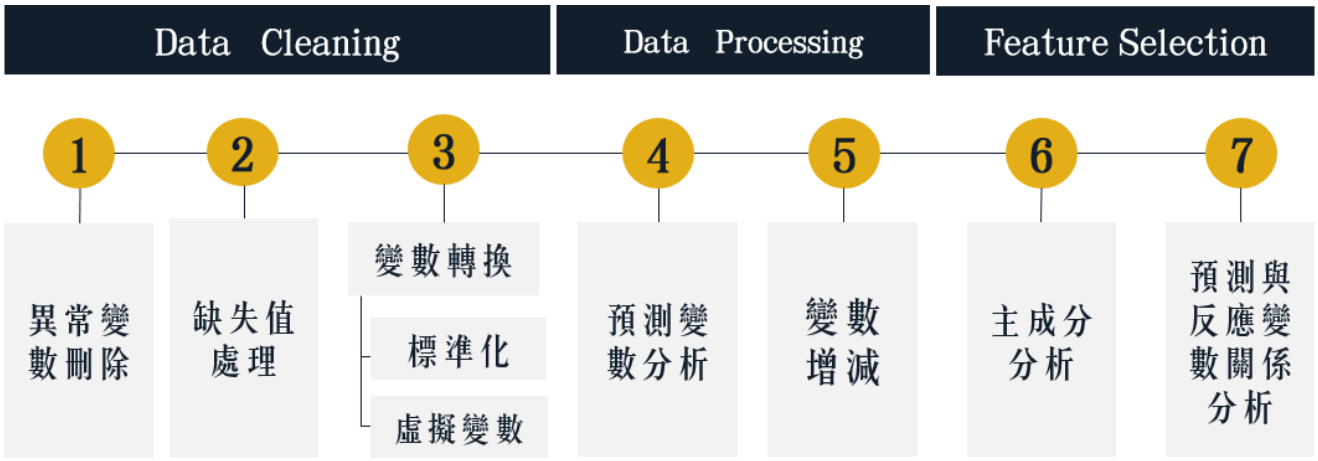
Cathay Big Data Competitions

目 錄

Table of Contents

一、研究方法 ---------- 3
(一)資料選取與特徵處理 ---------- 3
1. 資料清洗 ---------- 3
2. 資料處理 ---------- 6
3. 特徵選取 ---------- 7
(二) 模型選擇與成效驗證---------- 8
1. 模型選擇 ---------- 8
2. 成效驗證 ---------- 11
二、結 論 ---------- 12
三、參考資料 ---------- 14

（一）資料選取與特徵處理

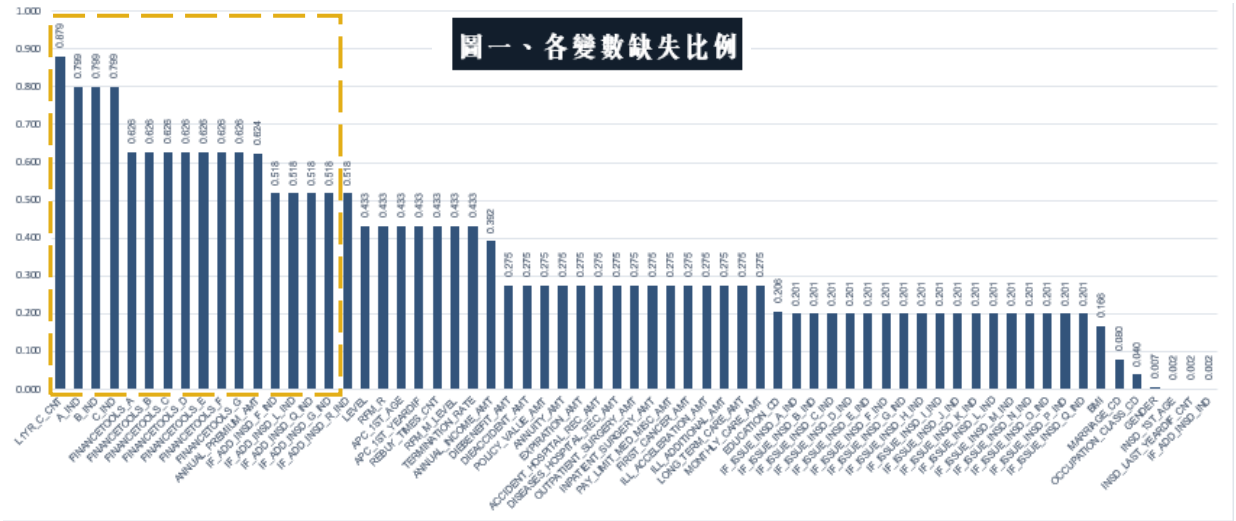


1. 資料清洗

資料含有缺失值的變量（如圖一）共 73 個。考慮缺失量的絕對及相對大小，變量類型及含義、缺失是否隨機、缺失變量之相關性、資料訊息的損失量及運行時間成本等因素，對缺失值進行刪除、合併及填補。

(1) 異常變數刪除（如表一）

圖一、各變數缺失比例



表一、刪除並處理缺失值過多之變數

變數	處理原因
L1YR_C_CNT	該變數的缺失比例達 88%，遠高於其他變數，所能提供的資訊極少，故直接將其刪除。

(2) 缺失值處理

(a) 產生 NA 變數（如表二）

表二、針對部分變數之缺失值產生 NA 變數

變數	處理原因
FINANCETOOLS_NA 未知使用任何理財工具	對於 FINANCETOOLS_A~G 缺失值多（如圖一），屬於非隨機缺失，不宜刪除。理財工具 A-G 之間存在顯著差異，且為聯動缺失，只要有一欄缺失，其餘六個變數皆缺失，因此新增 FINANCETOOLS_NA，即不願提及理財工具，有缺失值為 1，無則為 0，原變數的 NA 以 0 替代。
IF_ISSUE_INSD_I ND _NA 目前未使用任何 被保有效壽險保單	變數 IF_ISSUE_INSD_A~Q_IND 存在較多缺失且缺失聯動，各類別間有顯著差異,故使用類似做法，保留原始變數，新增變數 IF_ISSUE_INSD_IND_NA，即未知使用任何被保有效壽險保單，1 表示被保有效壽險保單有缺失值，0 表示無缺失，原變數的 NA 以 0 替代。

(b) 合併變數（如表三）

表三、針對部分變數進行合併

變數	處理原因
ABC_IND 有無訂閱電子報	A,B,C_IND 這類變數有近 80%的缺失值，爲了保留盡可能多的資訊，將其合併為一個變數 ABC_IND，即是否有訂閱電子報，0 為無訂閱，1 為有訂閱任何一種電子報，2 為未填答這三個變數。

(c) 決策樹方式補值

如圖一所呈現的之各變數缺失比例，部分變量已做處理,其餘變數以決策樹的方式進行補值。決策樹補值方法屬於多變量補值，利用除了反應變數 Y1 和流

水編號 CUS_ID 以外的變數預測含缺失值的變數。

由於在預測缺失值時未涉及反應變數 Y1，故可以將訓練集（Training set）和測試集（Testing set）合併，用更大的樣本資訊獲得更接近母體真實值的情況。在實際操作中，使用 rpart 包中 rpart()函數建構決策樹，這一函數允許選擇具有最小預測誤差的決策樹，即評估所有自變數和所有分割點，選擇組內數據的因變數取值變異較小的分割方式，默認尺度是 Gini 值，再使用 predict()函數對新數據進行預測，用預測值填補缺失。建模時，要區分類別變數與連續變數，分別對應「class」和「anova」的參數設定。此外，對類別變數做預測後需要將預測值進行類型轉換，從因子轉爲數值型再代入缺失空，以免出現資料類型不統一的錯誤。

(3) 變數轉換

將有序的類別變數視爲連續型變數，原始資料含有 45 個連續型變數和 86 個類別型預測變數，分類型對原始資料做變數轉換

(a) 連續型變數

連續型變數的單位不盡相同，下文的主成分分析對單位較爲敏感，需要做標準化及尺度化，消除單位的影響。在 45 個連續型變數中，我們僅針對其中 22 個未經神秘轉換和歸一化的變數做標準化，因爲並不清楚神秘轉換是否包含標準化。

(b) 離散型變數（如表四）

表四、多類別型變數轉換

變數	處理原因
L1YR_A_ISSUE_CNT	近一年沒有透過 A 通路投保新契約次數高達九萬多筆，因此選擇將投保次數轉換有無投保，1 代表有購買契約，0 代表無購買。
IF_ADD_INSD_IND CHARGE_CITY_CD MARRIAGE_CD	無順序的多類別變數需要先轉換成虛擬變數，放入模型才有解釋意義，避免距離歧義。

2. 資料處理

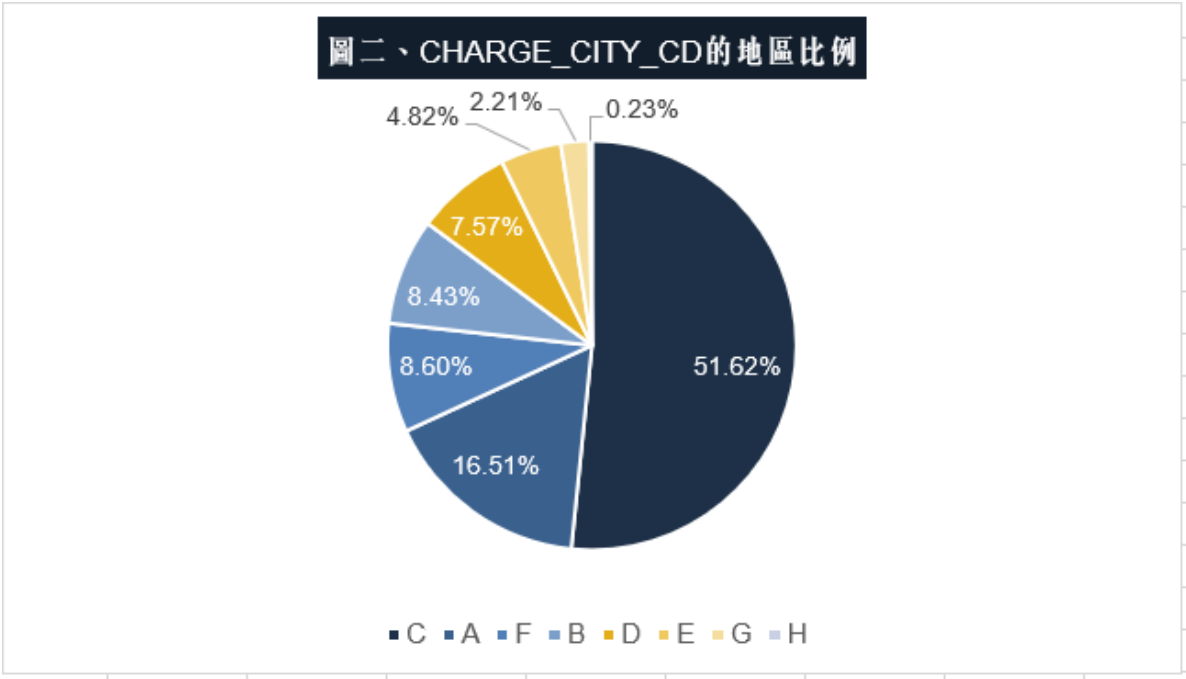
(1) 分析預測變數

由於各個變數間使用的單位不全相同，因此使用馬氏距離檢驗資料中的離群值。不用常見的歐式距離，而選擇用馬氏距離（Mahalanobis Distance）的原因在於馬氏距離有考慮到各變數之間的相關性，且不會受到尺度差異的影響。鑒於變數有類別型和連續型，我們僅選取所有連續型變數計算馬氏距離，並且將距離由大至小進行排列，將前 1%的資料集當作離群值予以刪除。

(2) 增減無缺失值變數（如圖二、表五）

表五、對 CHARGE_CITY_IS_C 增減變數

變數	處理原因
CHARGE_CITY_IS_C 收費地址是否是 C 地區	經 Kruskal-Wallis 檢定發現,不同區收費地址的反應變數 Y1 有明顯差異,再使用 Dunn 事後多重比較檢定分析兩兩間的關係可知,C 地區和其他地區存在顯著差異,此外,收費地址為地區 C 的樣本比例達 51%(如圖二),因此創造新變數 CHARGE_CITY_IS_C,1 為收費地址 C 區,0 表示收費地址是其他地區。



3. 特徵選取

(1) **主成分分析**（Principal Component Analysis, PCA）（如表六）

表六、對相關性高之變數進行主成分分析

變數	處理原因
IM _pca_PC1 IM _pca_PC2	IM_CNT 特定商品持有類別數以及 IM_IS_A,B,C,D_IND 此 5 個變數彼此相關係數很高且具相似的資訊，因此利用主成分分析萃取出前兩個解釋比例達 91%的主成分作為新的變數。
Hospital_pca	由於 ACCIDENT_HOSPITAL_REC_AMT、DISEASES_HOSPITAL_REC_AMT、OUTPATIENT_SURGERY_AMT 以及 INPATIENT_SURGERY_AMT 這 4 個變數相關性高且屬相同類型的變數，因此利用主成分分析萃取出第一主成分，解釋能力約 90%。
IF_ADD_X_IND_pca_PC1 IF_ADD_X_IND_pca_PC2 IF_ADD_X_IND_pca_PC3	IF_ADD_F, L ,Q,G,R_IND 此 6 個變數之間的相關係數很高而且變數代表著相似的資訊，因此利用主成分分析的方式萃取出前三個解釋比例達 92%的主成分作為新的變數。

(2) 預測變數與反應變數之關聯分析（如表七）

這部份我們透過 ANOVA、Kruskal-Wallis test 及卡方獨立性檢定，檢定樣本在不同類別下，對反應變數是否有差異。通常檢定 3 組以上的平均值是否相等時會使用 ANOVA，不過需要變數符合常態分配的基本假設與變異數同質性檢定，因此我們會傾向使用 Kruskal-Wallis test 或卡方獨立性檢定等無母數方法，以檢驗各組平均等級的差異，若檢定的結果為拒絕虛無假設，代表各組的平均值應該是相等，便會考慮刪除該變數。也透過相關分析，尋找變數之間是否有共線性的類別，將相關性較低之類別刪除。

表七、預測變數分析之方法及處理原因

變數	方法	處理原因
IF_ADD_INSD_ F, L, Q, G, R_IND	敘述統計	目前是否壽險保單被保有效類別共 5 類，從反應變數可發現，若只針對有購買重疾險之客戶分析，僅一位客戶是無任何壽險保單，因此我們認為 IF_ADD_INSD_IND 是否投保附約（被保）就已包含相關資訊，因此刪除這 5 個變數。
DIEACCIDENT_AMT	相關分析	不論是補值前後 DIEACCIDENT_AMT 和 DIEBENEFIT_AMT 的相關係數皆大於 0.9，又考慮到兩變數與反應變數之相關，選擇將其刪除。
CONTACT_CITY_CD	ANOVA	在不同縣市的聯絡地址下，無足夠證據說明反應變數有差異，因此推論該變數和反應變數無關並刪除。
CHANNEL_B_POL_ CNT	Kruskal-Wallis test	透過檢定發現，無足夠證據說明在反應變數不同的情況下有差異，因此推論該變數與反應變數較無關並刪除。
IF_Y_REAL_IND	卡方獨立性檢定	透過檢定發現，無足夠證據說明和反應變數相關，因此決定將其刪除。

（二）模型選擇與驗證成效

1. 模型選擇

(1) **XGBoost（eXtreme Gradient Boosting）**

以梯度提升決策樹（Gradient Boosted Decision Tree）為基礎改良的算法。GBDT 以集成學習（Ensemble Learning）的技術，將多個決策樹疊加，將先前決策樹保留不變，再建構新的預測模型，並針對每個決策樹調整參數（權重），進而優化過去決策樹的效能。XGBoost 與傳統 GBDT 最大的差別在

於，新增決策樹主要是利用網格搜尋（Grid Search）找到最佳的決策樹。在目標函數裡面，除了傳統機器學習或是統計學上的損失函數，增加了一個懲罰項，針對過於複雜的模型給予扣分。且有許多優化其計算速度的資源，因此其為我們主要的預測模型。

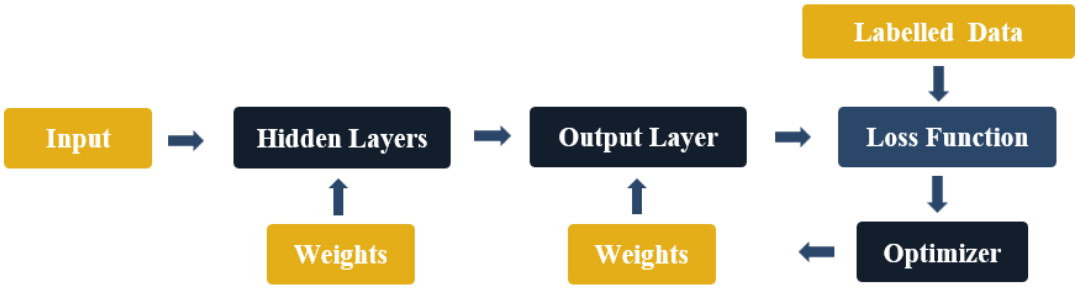
(2) 邏輯斯迴歸分析（Logistic Regression）

以迴歸分析為基礎，針對二元反應變量所衍生出的統計方法。傳統迴歸分析通常為針對反應變量的定義域為實數空間，抑或是一段連續性的實數集合。因此面對反應變量為布林值的問題不適合使用。邏輯斯迴歸分析的精神亦與傳統迴歸分析不同，傳統迴歸分析是期望能找到一條空間上的直線，讓所有資料點愈貼近該線愈好。此法是期望能透過模型計算找到一個臨界點，將資料分為「較可能發生」以及「較不可能發生」兩群。

(3) 類神經網路（Neural Network）

此模型（如圖三）的好壞取決於我們採用何種損失函數、成效衡量指標（Metrics）、激勵函數（Activation Function）、優化器（Optimizer）或隱藏層（Layers）等，需要經反覆實驗才能得到較好的模型，使預測更為準確。在本例中，根據二元反應變數設定損失函數為 binary crossentropy；輸入層和隱藏層之激勵函數採用 relu，排除負值，隱藏層到輸出層使用 sigmoid，求出預測機率並設定成效衡量指標為 precision，計算所有「正確被分類的結果」佔所有「實際被分類到的」的比例；優化器選擇 Adam，並設定較小的學習率。設定 batch size 分批求解加速梯度下降的過程。

圖三、Neural Network 流程圖



(4) Staking（Stacked Generalization）

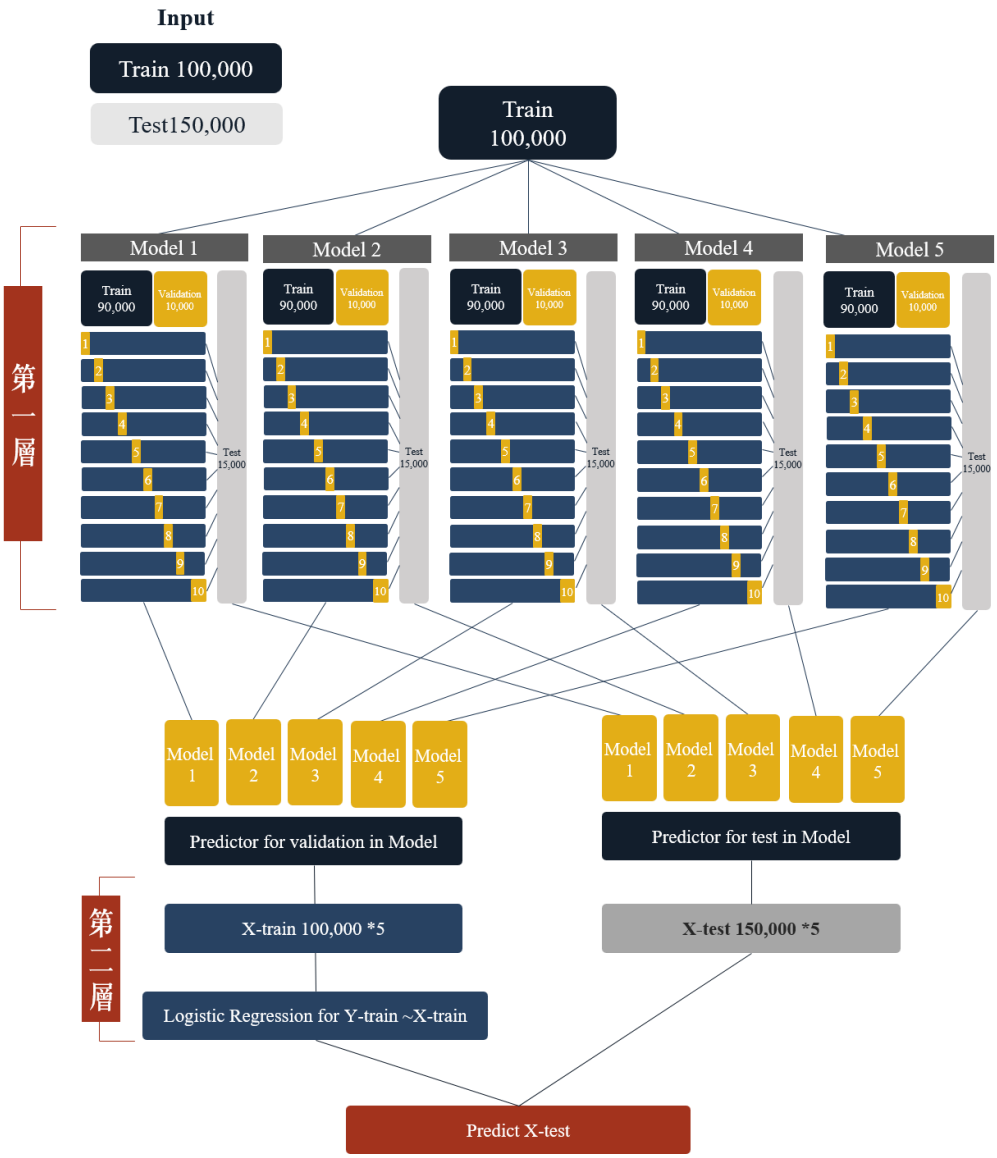
利用 Stacking 技巧將不同的資料前處理、不同的預測模型的產出結合。亦是屬於集成技巧。首先固定一個模型以及其超參數，接著將訓練資料集分層切分成十等份，保留 Y1 = 0 和 Y1 = 1 的比例關係，將資料隨機切分成等份。

接著依序抽出一份作為「驗證資料集」(Validation Set)、另外九份為「訓練資料集」，利用訓練集來訓練模型，再對驗證集以及測試集做預測，重複十次。

結束後產出驗證集預測值（為整份訓練集）以及測試集的預測值。再以不同模型重複上述運算，得到多個模型的驗證集預測值以及 10 個測試集的預測值。由於驗證集我們已知反應變數，因此我們可以透過驗證集觀察到，這些模型預測值跟真實反應變數之間的關係。再利用模型，建立出驗證資料集以及其與反應變數的關係。最後，套用至 10 個測試集預測值平均，得到最後的測試集預測值。

比較「單用 XGBoost」及「Stacking 技巧及單用一個 XGBoost 模型」後發現，用 10 個測試資料集預測值平均所做出的 AUC 表現會較高，因此我們亦會利用 Stacking 技巧及單用一個 XGBoost 模型來進行預測，流程圖如圖四。

圖四、Staking流程圖



2. **模型驗證**

模型	AUC 平均數	AUC 標準差	模型結果
XGBoost	0.8359	0.0090	XGBoost 相較其他決策模型，如隨機森林，皆為集成模型，因此其模型表現也較佳。因有許多算法優化的資源，其計算速度不亞於其他 AUC 更低的方法。
利用 Stacking 技巧及單用一個 XGBoost 模型	0.8428	0.0099	其模型平均 AUC 較單純使用 XGBoost 來得好，因此我們會傾向採用此技巧，而非單純使用 XGBoost 來預測。
利用 Stacking 技巧，第一層為兩個 XGBoost 、第二層為 Logistic Regression	0.8444	0.0104	其平均值高於單用一個 XGBoost 模型所做出來的 AUC，代表 Stacking 擁有提高模型表現的效果。由於 Stacking 會分層切分資料，而切分資料會影響模型的表現，因此其標準差會比上述兩個算法來得高一些，不過上升幅度仍在我們可以接受的範圍內。
利用 Stacking 技巧，第一層為兩個 XGBoost 、第二層為 Neural Nework	0.8487	---	該模型相較於邏輯斯迴歸會考慮非線性關係，且神經網路是一種自適應系統，具備學習功能；多種模型的組合可以增加預測的穩健性。

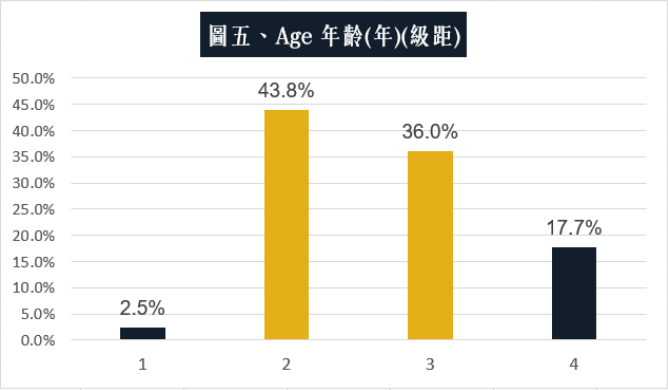
結 論

透過「利用 Stacking 技巧，第一層為兩個 XGBoost、第二層為 Neural Nework」之模型，可以選出五個變量，其為影響是否購買重疾險重要因素。分別為年齡(Age)、客戶職業類別對核保風險程度(OCCUPATION_CLASS_CD)、往來關係等級(LEVEL)、最近一次被保人身份投保距今間隔時間(INSD_LAST_YEARDIF_CNT)及當年度保障_一般身故(DIEBENEFIT_AMT)。

接著再以預測「會購買重疾險」的機率值 0.0166 作為閾值，只要超過閾值便會視其為會購買重疾險保單，反之則不會購買。經結果發現有 37,181 位客戶預測會購買重疾險保單（佔測試集中的 2.5%），以下我們將從這些客戶中，針對上述五項重要因素做說明。

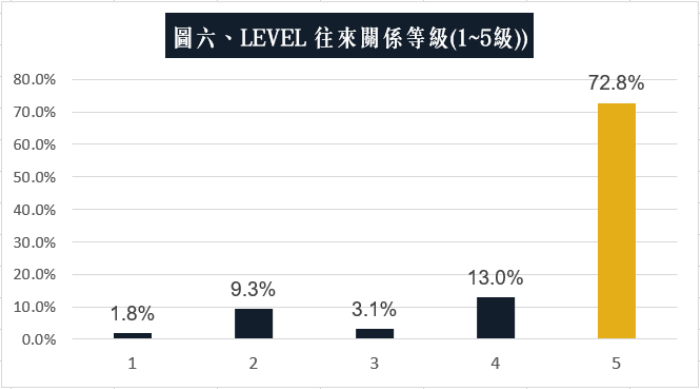
1. 年齡(Age)

從圖五可以發現，年齡級距為 2 者(43.8%)，有較高的機會購買重疾險保單，其次是年齡級距為 3 者(36%)。然而，年齡級距為 1 者(2.5%)，購買重疾險保單之機率較低。因此推論年齡是會購買重疾險保單的重要因素。



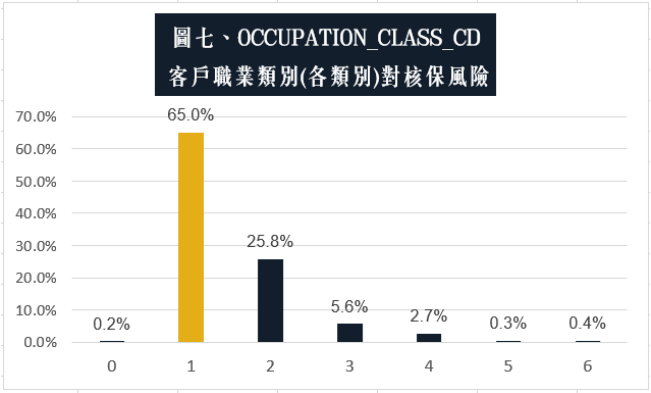
2. 往來關係等級(LEVEL)

從圖六可以發現，往來關係等級為 5 者(72.8%)，有較高的機會購買重疾險保單。然而，其他 1 至 4 級往來關係等級，購買重疾險保單之機率顯著較低。因此推論往來關係等級是會購買重疾險保單的重要因素。



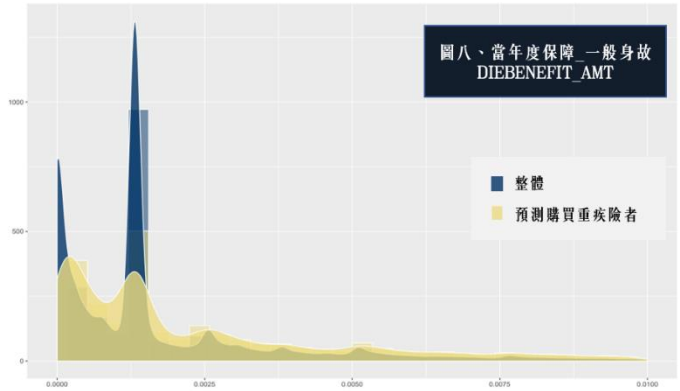
3. 客戶職業類別對核保風險程度(OCCUPATION_CLASS_CD)

從圖七可以發現，客戶職業類別對核保風險為 1 者(65.0%)，有較高的機會購買重疾險保單，其次是客戶職業類別對核保風險為 2 者(25.8%)。然而，其他類客戶職業類別對核保風險，購買重疾險保單之機率顯著較低。



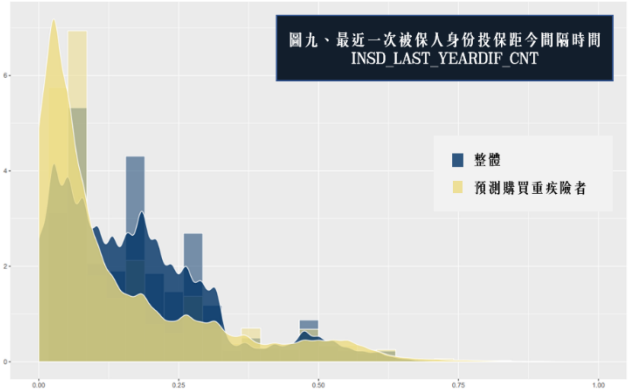
4. 當年度保障_一般身故(DIEBENEFIT_AMT)

透過 Kruskal-Wallis test 檢定結果發現，若該客戶（被保人身分）當年度保障_一般身故之保障金越高（p-value < 0.001），會有較高的機會購買重疾險，從圖八的分布則可以發現值與整體相比，預測購買重疾險者其一般身故之保障金大部分介在 0.0000 至 0.0025 之間。



5. 最近一次被保人身份投保距今間隔時間(INSD_LAST_YEARDIF_CNT)

透過 Kruskal-Wallis test 檢定結果發現，最近一次被保人身份投保距今間隔時間越短，會有較高的機會購買重疾險（p-value < 0.001），從圖九的分布則可以發現值與整體相比，預測購買重疾險者其投保距今間隔時間大部分介在 0.00 至 0.125 之間。



綜上所述，可透過我們的模型預測目標客戶是否會購買重疾險，亦可依據這五項重要變數，初步篩選出未來推薦重疾險保單之主要對象。

參考資料

1. 缺失值處理之方法

Analytics Vidhya Content Team (2016, March 4). Tutorial on 5 Powerful R Packages used for imputing missing values. *Analytics Vidhya*. Retrieved October 4, 2019, from https://www.jamleecute.com/missing-value-treatment-%E9%81%BA%E5%A4%B1%E5%80%BC%E8%99%95%E7%90%86/

2. 馬式距離之定義與應用

Stephanie (2017, November 21). Mahalanobis Distance: Simple Definition, Examples. *Statistics How To*. Retrieved October 4, 2019, from https://www.statisticshowto.datasciencecentral.com/mahalanobis-distance/

3. Kruskal–Wallis One-Way Analysis of Variance

Zar, J.H. (2010). Biostatistical Analysis. Fifth Edition. Prentice-Hall, Upper Saddle River, NJ, Chapter 13.

4. XGBoost 模型之定義與應用

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

5. 邏輯斯迴歸分析之定義與應用

Tommy Huang (2018, September 26). 機器/統計學習: 羅吉斯回歸(Logistic regression). *Medium*. Retrieved October 5, 2019, from https://reurl.cc/k5EdNx

6. 類神經網路

Mike Chen (2017, December 13). Day 03：Neural Network 的概念探討. *Medium*. Retrieved October 5, 2019, from https://ithelp.ithome.com.tw/articles/10191528