第三屆商業模式與大數據分析競賽

人工智慧金融挑戰賽
競賽計畫書

中華民國 109 年 10 月 29 日

競賽摘要

競賽主題	預測資產變化趨勢，發掘潛在需求
競賽摘要	我們根據 2019/5 至 2020/5 月份資料，建構 XGBoost 模型來進行 2020/6 資產總額的預測。以第一到第十二個月的特徵變數與第十三個月的資產總額建立模型，並使用第二到第十三個月的特徵變數預測第十四個月的資產總額，使用預測出的結果計算均方誤差(MSE)評估模型的表現。 在預測第十四個月的總資產後，為了瞭解各種不同特質的客戶對於各項金融商品的需求，我們先使用上採樣的方式重新合成樣本以改善資料不平衡的問題，再針對所合成的樣本使用決策樹進行分析，幫助我們釐清年齡、性別及職業類別如何影響客戶對於金融產品的需求，我們發現年齡是影響客戶是否購買產品的重要因素，而其中年紀較輕者其購物行為可再取決於職業類別，年紀較長者之購買意願則會取決於性別。在知曉客戶個人狀態與對金融產品的需求之後，我們期望能依據產品本身的性質與客戶的特質設計出適合各項產品的商業模式，而我們所提出的三個策略分別是—購買產品 I 減免手續費的方式來穩固客群；凡購買產品 J、K 即享有低利率的貸款方案以吸引新的客源；透過交易往來頻繁的 G 通路為產品 L、M 做宣傳，提升產品 L 及 M 的曝光率。

目錄

一、緒論

現今的銀行提供的金融服務，已經不同於以往。比起親自前往臨櫃找個理財專員，去操作投資決策，現今只要打開手機，用 App 就可以直接下單，甚至還能透過目前市場各類訊息，自動推薦投資組合。以上的服務皆不受銀行在下午三點因內部清點而關門的限制，皆能透過網路及手機就能完成。

過去銀行的市占率是透過實體據點，也就是分行數量來取勝，換句話說，能有最多機會接觸客戶，自然能有不低的市占率。但隨著客戶使用習慣的改變，實體銀行需要逐漸轉型，與其他銀行做出差異。未來必須建立通路整合的管理機制及洞悉顧客的行為模式，強化銀行與客戶端的連結，創造嶄新的金融服務體驗，才能在競爭激烈的市場中克敵制勝。

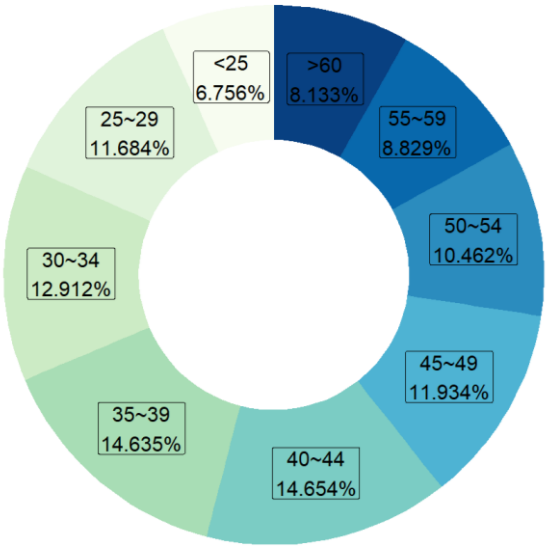
本分析案中，銀行覆蓋客群職業類型眾多，我們想透過不同職業之客戶對於金融服務的使用通路、貸款商品的餘額、甚至是目前已申購哪些理財商品等等指標，去預測客戶未來去購買對於本行金融商品(產品 I~M)的意願，甚至預測該客戶在這些商品中的資產總額。

此外，我們也將透過機器學習的方法去探討影響客戶持有各類金融商品的資產總額的因素，利用分析結果來訂定未來推廣金融商品所需要的策略。例如:發現 A 商品往往都是在特定年齡層的客戶才有購買跡象、或是習慣使用特定通路交易往來的客戶才有較高比例來購買，如果想要推廣此類產品，可以在欲推廣的新型通路上推出折扣優惠、或是對於不同年齡層的客戶制定專屬優惠等等。

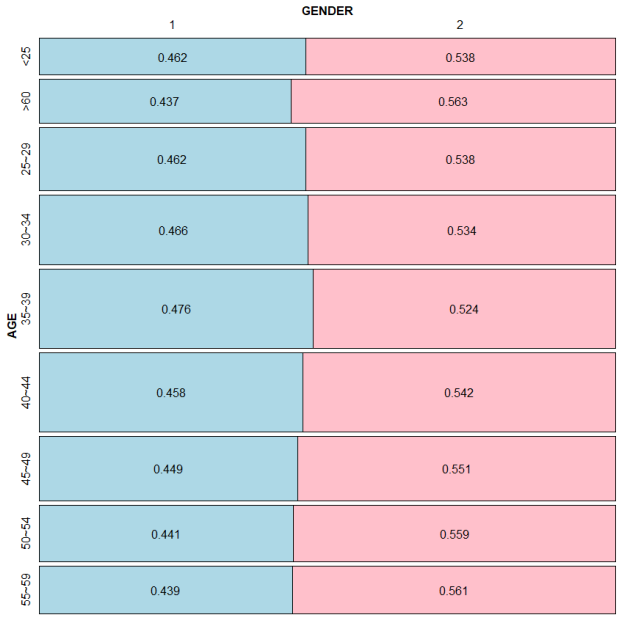
總合上述，我們期望透過洞察數據來發現潛在資訊並開創新的商業模式，期望能促使客戶獲得有效且即時的資源及服務，藉此提升客戶端的滿意度以及吸引新的客源。

二、探索性資料分析

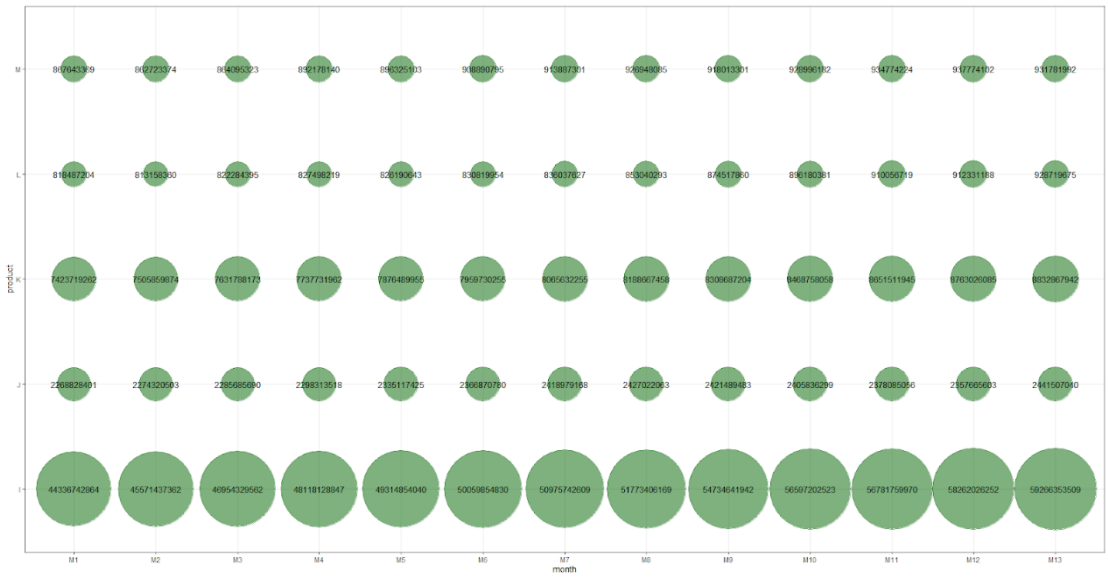
(1) 客戶年齡層比例，可以觀察出客戶年齡層落在 35~45 歲較多，小於 25 歲以及 60 歲以上占比較少。



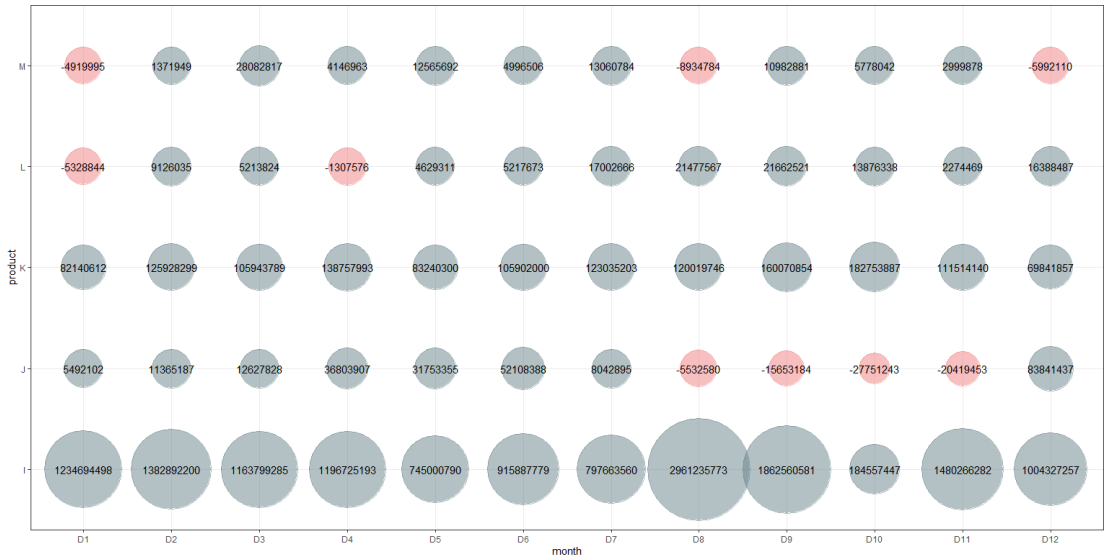
(2) 觀察各齡層的性別比例，同時也可從寬度觀察各年齡層人數多寡。各年齡層的性別占比大致上是平衡的。



(3) 探討各月份持有各資產的金額變化，橫軸為月份:2019/5 至 2020/5(由符號 M1~M13 表示)，縱軸為各產品(I/J/K/L/M)，圓形大小為各月份各資產持有金額。可以發現客戶持有產品 I 的金額明顯較其餘產品多。

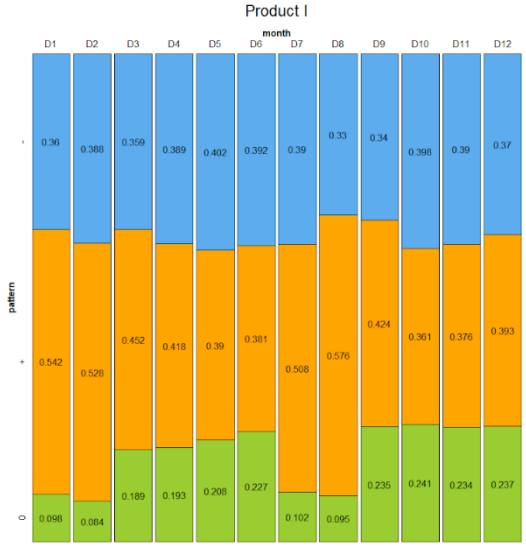


(4) 探討各月份各資產的持有"金額差異"變化，橫軸為前後一個月份金額差異(ie. D1=M2-M1)，縱軸為各產品，圓形大小為各月份持有各資產金額差異量，依顏色區分正負值。可以觀察到客戶持有產品 I、J、K、L、M 的金額在各月份的增減狀況。全部產品中，客戶持有產品 I、K 的資產金額持續成長；客戶持有產品 I 的資產金額在 2019/8 成長最多。客戶持有產品 J 的資產金額在 2019/8、9、10、11，四個月持續遞減。

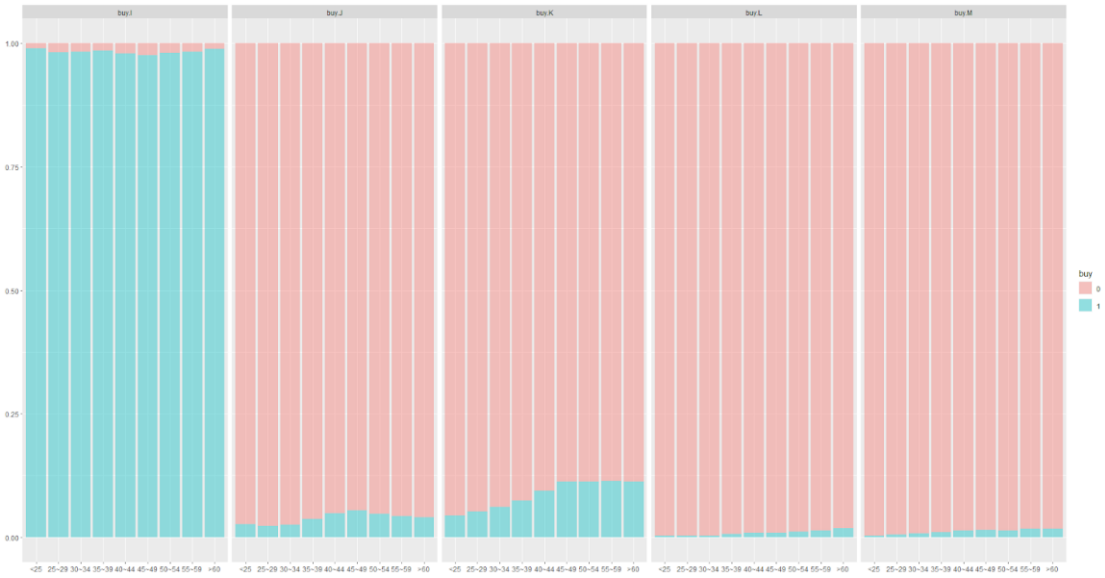


(5) 探討不同產品不同月份間金額變動趨勢的客戶比例。

觀察前後月份持有不同產品的資產金額差異量趨勢的客戶比例圖，縱軸為當月份資產金額差異(D1=M2-M1)；橫軸為當月資產金額變化趨勢(負號代表減少投資金額; 正號代表增加投資金額 ;0 代表不變)的客戶比例。在 2019/1、2、7、8 該月份間，客戶增加投資產品 I 的客戶比例明顯增加，而不變的客戶比例明顯減少，客戶減少投資產品 I 的客戶比例大致都落在 0.3~0.4 之間。

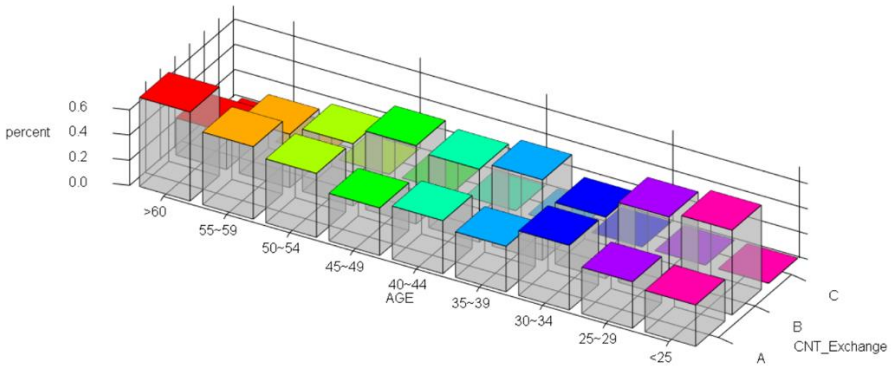


(6) 觀察各年齡層有無購買 I/J/K/L/M 產品的分布狀況，可以發現 13 個月中曾購買產品 K 的年齡層分布較多落在年齡層 40 歲以上

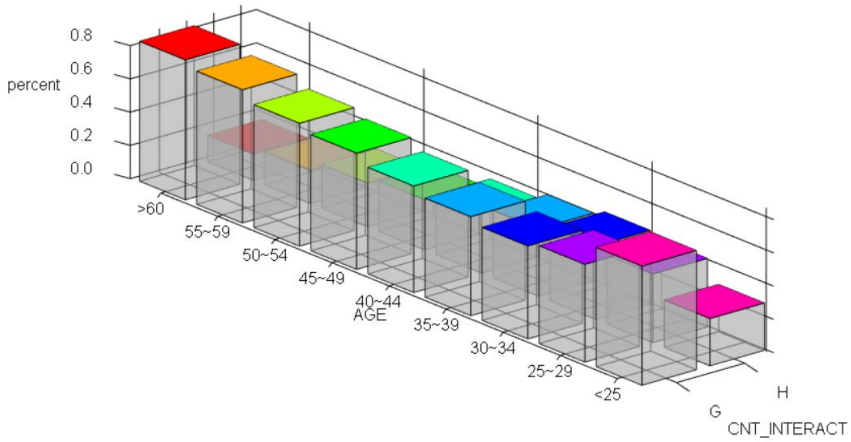


(7) 觀察各性別有無購買 I/J/K/L/M 產品的分布狀況，可以發現不同性別有購買產品 I 的比例差異不大，性別 2 會購買產品 J,K 的比例較性別 1 多。

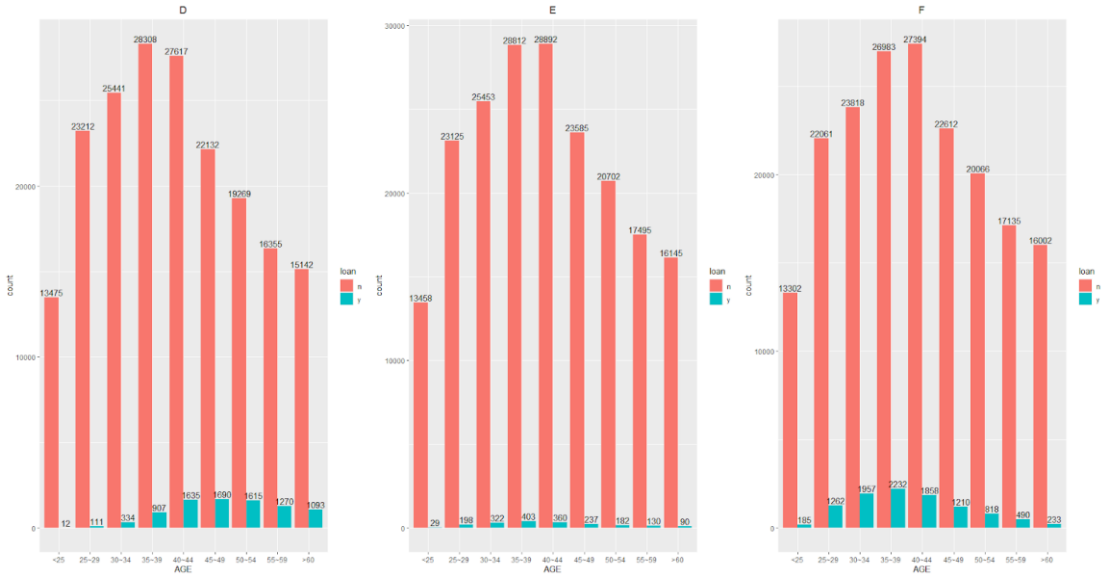
(8) 探討各年齡層中，不同通路換匯的比例，觀察不同年齡層是否有偏好的換匯通路。年齡 55 歲以上的人偏好使用通路 A 換匯；45~35 歲以及 25 歲以下的人偏好使用通路 B 換匯, 50~54 歲以及 30~34 歲，使用通路 A,B 換匯的比例差不多。



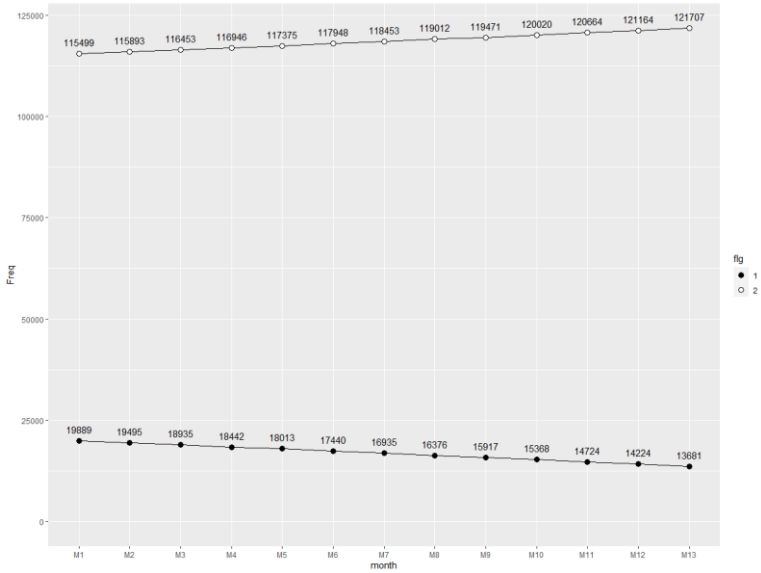
(9) 探討各年齡層中，不同通路往來的比例，觀察不同年齡層是否有偏好的往來通路。可以發現各年齡層皆偏好通路 G；而從通路 H 的角度來看，25~35 歲年齡層往來比例較其他年齡層高。



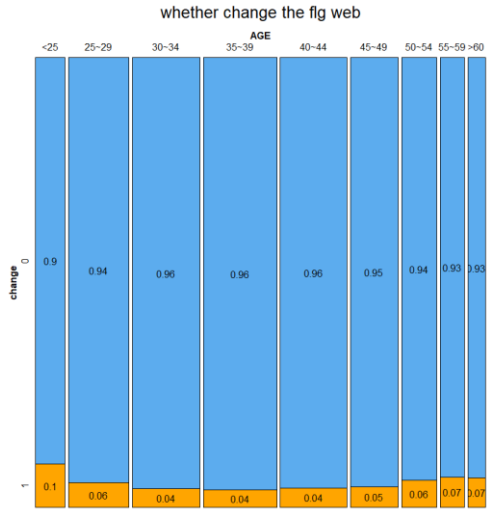
(10) 觀察不同貸款類別，在各年齡層有無貸款的狀況。可以發現客戶有貸款的人占少數，有貸款類別 D 的年齡層較高；有貸款類別 E 的客戶最少。



(11) 觀察不同客戶身分類別隨著月份的變化，可以發現數位會員標籤 1 隨著月分減少，數位會員標籤 2 隨著月份增加。



(12) 觀察 13 個月中各年齡是否變動會員身分(數位會員標籤 1 變為數位會員標籤 2)的比例，可以發現年齡層小於 25 歲有更動會員標籤的比例最高。



三、特徵資料處理過程

(1) 遺失值處理

年齡屬於類別型變數，共有 474 筆資料為遺失值，我們使用 K-Nearest Neighbours(KNN)的方式來進行插補，藉由使用持有資產 I~J 金額的資訊來判斷哪些年齡的人與缺值的人有相似的情況。數位會員的標籤則有 64704 筆的遺失值，數位會員標籤的 1 和 2 代表客戶會員身分的不同類別，能使用的服務不同，因此我們選擇將遺失值視為新的一類，代表不同於類別 1 與類別 2 的人。對於客戶的工作變數則有高達 47.6%的遺失值，我們認為遺失的數量過大，即便進行插補也沒有意義，因此合理的處理方式為刪除變數。

(2) 類別變數處理

對於類別變數，會轉為數值型變數還再做進一步的分析，我們使用的方式為 One-hot Encoding 的方式將類別變數轉換為 dummy variable，例如客戶所屬地區共有 10 個地區，則會增加相對應數量的新變數，對於客戶的每個例項只有其對應的類別屬性為 1，其他則會是 0。

(3) 異常值處理

由於特徵變數的名稱有經過去識別化的處理，不易理解其變數所代表的意義，且金融市場與資產的變化波動大，因此難以界定哪些值屬於異常值，最終我們選擇不處理異常值。

(4) 特徵變數的轉換

我們所使用的轉換方式有標準化(standard scaling)與歸一化(min-max scaling)兩種，在我們經過後續模型的嘗試之後可以發現經過兩種轉換後我們所得到的結果都不如使用原始資料的結果，因此我們使用原始資料特徵來進行分析。

(5) 新增特徵變數

a. 經過觀察後可以發現對於紀錄次數的變數，即換匯次數和往來次數，這些變數又會區分成不同的通路來記錄，例如換匯次數當中有通路 A 到 C，而往來次數有通路 G 到 H，但大多數的值高達 7 成以上都為 0，我們認為由於 0 的數量過多，對於預測客戶次月的資產金額，區分不同通路可能沒有意義，因此考慮新增變數，將所有通路的值相加起來，例如新增變數 CNT_Exchange_SUM_M1，即代表在 2019 年 5 月份時，通路 A 到 C 的換匯次數總和。最終我們有新增 CNT_Exchange_SUM 與 CNT_INTERACT_SUM 的 M1~M13 等變數。

b. 由於後續的模型建構之中，有採取無法考量時間資訊的模型，而又為了掌握月份與月份間變化的趨勢，我們參考了「滑窗法」的概念。滑窗法為重新組織數據，創立新的變數，使其成為能應用時間序列資料於監督式學習的一個方法。具體的方法為通過前一個時間的值來預測下一個時間的值，例如若我們運用第 13 個月份來建構模型，則會使用第 12 個月份來作為特徵變數，因此我們新增了變數－shift_var。

四、模型的建構、預測與結果分析

(1) 構思

我們的目標是藉由 2019 年五月到 2020 年五月，總共 13 個月的資料來預測 2020 年六月份的資產總額。起初，考慮到有明顯時間序列的問題，因此我們初步嘗試使用 RNN 與 LSTM 兩種深度學習的方法來進行建模與預測，但最終我們無論如何抉擇 feature 放入與捨棄或是參數的調整，最終得到的結果都會發現所有客戶的資產預測結果都明顯不優。因此我們推論可能「時間」的因素在這筆資料下，可能效應沒有非常明顯。接著，我們選擇使用機器學習的 XGBoost 方法來進行預測，我們同時結合多元時間序列分析中的滑窗法的概念詳細說明如第三節，來使模型學習出時間趨勢的變化，但由於先前的討論，我們僅使用 lag(1)。詳細的建構說明如下段說明。

(2) 步驟

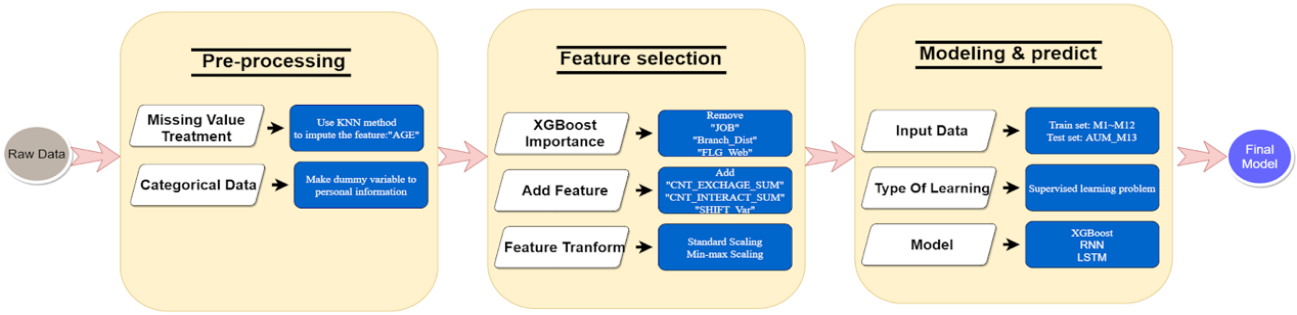
首先是進行特徵值前處理，而處理方式的詳細說明如第三節，主要順序為遺失值處理，類別變數轉換，新增特徵變數，接著是依據模型的 importance 來刪除不重要的變數，最後留下重要變數後建立模型，並調整參數，最後得到最終模型。我們使用均方誤差(MSE)來做為衡量模型的方式，若誤差數值越低則代表模型表現較好。

(3) 建模與結果

首先，我們先使用第一到第十一個月的特徵變數與第十二個月的資產總額來建立模型，並使用第二到第十二個月的特徵變數來預測第十三個月的資產總額，並且觀察預測出的結果來計算 MSE 來觀察模型的表現，我們可以得到 $4.701*10^{11}$，我們認為模型有不錯的表現，因此認為這種模型建構方式是可使用的。最終我們會使用第一到第十二個月的特徵變數，與第十三個月的資產總額來建構模型，最後使用第二到第十二個月的特徵變數來預測第十四個月的資產，也就是 2020 年六月的客戶資產總額。

(4) 流程圖

我們的模型建構過程如下圖:



(5) 預測結果

透過觀察後可以發現多數人的資產總額通常不會有太過巨大的變化，而由於我們沒有第十四個月的資產金額，因此可以與上一個月的資產總額進行比較，部分客戶預測結果如下:

客戶編號	2020 年 5 月	預測 2020 年 6 月
161236	12857	12741
163040	364722	360441
68474	1111970	1093610
159801	40708	39993
117899	463309	457224
155876	460582	461413

由於我們的模型預測出的資產金額可能會有負數，可能原因為該客戶的資產一直都過小所導致，因此最終我們的預測結果會將負數的預測值改為 0，而小數點後的值則四捨五入進位。

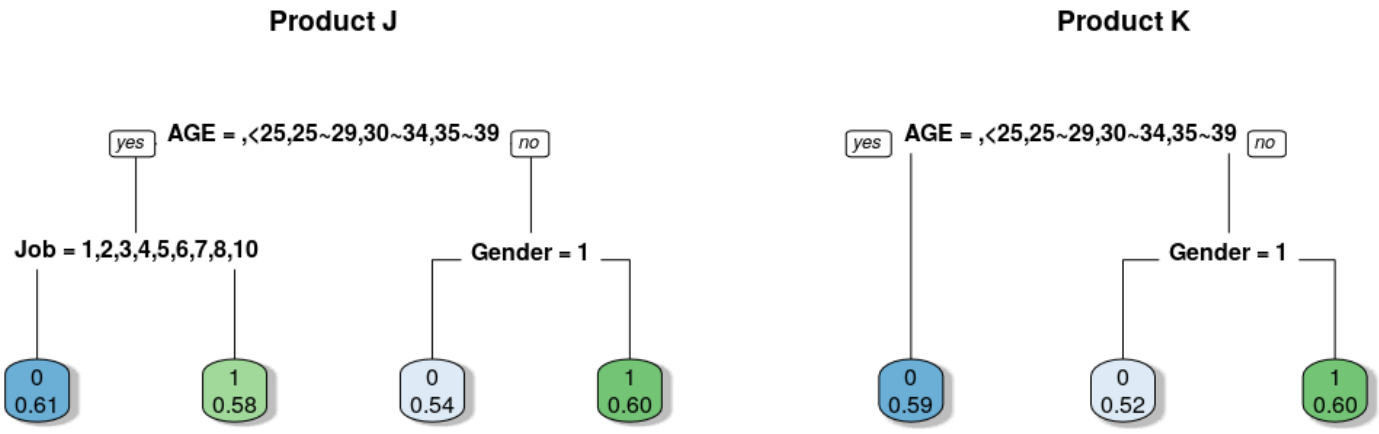
五、客戶人生階段和金融商品需求的邏輯性及關聯性

為了瞭解各金融商品的購買情形，將客戶在 2019/5 月到 2020/5 月之間有購買商品 I、J、K、L、M 紀錄的比例統整如下表，表中可以看到商品 I 的購買比率極高，資料當中幾乎所有人都有購買該產品，因此推斷此一商品可能為入門門檻低或是必要性高的產品，客戶不論處於何種狀態應當都有對此產品的需求；而商品 L,M 的購買比率則非常低，我們推斷這兩類商品的本質促使客戶不論處於何種狀態都沒有什麼購買意願，因此接下來將針對金融商品 J,K 進行更進一步的探討。

產品	I	J	K	L	M
購買比例	98.3%	3.9%	8.6%	0.8%	1.2%

我們想藉由客戶的年齡、性別及職業這些基本資料來探討購買金融商品 J,K 的客戶有何種特徵，以推論客戶人生階段和金融商品需求的邏輯性及關聯性，然而，即便購買 J,K 商品的人分別有 3.9%及 8.6%，仍有 90%以上的客戶並未購買此二商品，資料存在不平衡的問題使得若我們直接使用原始資料找尋潛在關係會得到偏頗而沒有意義的結論，為了解決此一問題，我們採用上採樣的方式重新合成適合做後續分析的樣本。

藉著重新合成的資料，我們使用決策樹幫助我們了解年齡、性別及職業對於是否曾經購買金融產品的影響。決策樹的結果如下圖所示，圖中最底下的葉節點 0 表示在 2019/5 月到 2020/5 月之間沒有購買該產品、1 則代表 2019/5 月到 2020/5 月之間有購買該產品的紀錄，而下圖（左）為 J 產品的決策樹、下圖（右）則為 K 產品的決策樹。



產品 J 的決策樹顯示，若顧客年齡落在 40 歲以下，則購買此產品與否取決於職業總類—職業類別為 9 及 11 的客戶有較高的意願購買 J 產品；若顧客年齡落在 40 歲以上，則購買 J 產品與否取決於客戶性別—客戶性別為 2 時有較高的意願購買 J 產品。產品 K 的決策樹亦有異曲同工之妙，年齡同樣是決定客戶是否購買此產品的重要因素，通常年紀 40 歲以下的客戶並沒有購買 K 產品的意願，而年齡 40 歲以上者購買行為的不同則可歸因於性別。

透過這樣的發現，我們可以歸納出年齡、性別及職業類別對於金融商品需求的關聯性—以常理而言，年齡是影響客戶本身積蓄與投資能力的指標，年紀較輕者通常沒有太多餘裕可以對非必要的金融產品進行購買或投資，除非其職業較為特殊能夠支撐此一開銷才會促使客戶有投資意願，因此年紀輕者通常傾向只購買產品 I，是否購買其他產品則可以透過職業類別進行推斷；對於年齡大於 40 歲的客戶，他們通常已經累積了一定的財富，其購買行為則容易受性別影響。其中，年齡分界點透過這筆資料可以訂定在 40 歲，40 歲以下及以上的客戶對於金融產品有不同的需求和考量。

六、行銷金融商品的商業模式

對於不同性質的產品，理應根據其特性建立不同的商業模式，才能夠完善運用行銷資源，最大化公司的獲利。因此，在探討如何對金融商品設計適當的商業模式時，我們將依據資料所提供之資訊對產品依其特質分類，對於性質相似的產品訂定適當的銷售手法，期望能藉此提升目標客戶的購買意願。從資料中我們歸納出五種金融商品可以分為三類，分別為幾乎所有人都有購買的 I 產品、特定年齡層有較高購買意願的產品 J 及 K、整筆資料中僅有約 1%客戶購買的產品 L 及 M

(1) 金融產品 I

由於金融商品 I 在這筆資料當中有非常高的購買率，客戶不分任何類型幾乎都有購買此一產品，因此推斷產品 I 應當是入門門檻低或是必要性高的產品。我們相信這樣類型的商品不管在哪間銀行應當都有銷售，而若要讓客戶持續在本行購買此商品，需建立顧客的忠誠度。為了達到此目的，我們認為在行銷 I 產品時，可以提出提升顧客交易方便性的方案，像是若客戶購買 I 產品則給予他們「在任何與銀行的交易時有手續費減免」的優惠，透過建立熟客制度，讓客戶願意在未來持續於本行購買此商品。

(2) 金融產品 J、K

金融商品 J 及 K，購買率雖低，但特定年齡層仍有一定的購買比率，所以假定為購買門檻較高，需要有較穩定的經濟能力才能支付的商品。歸納購買此類產品的客戶分布，持有率佔據前三高的客群為大於 40 歲的壯年人口及中年人口，且透過探索性資料分析結果可以推斷，大於 40 歲的客群多數都有車貸或房貸的負擔。為了持續穩固這年齡層的客群，我們可以推出凡是購買 J、K 產品，則「透過本行辦理貸款，如信貸、房貸、其他貸款等，能享有更低的循環年度利率」。

(3) 金融產品 L、M

在這筆資料的中，僅有約 1%的人有購買產品 L 或 M，我們推測這兩類產品除了入門門檻高外，它們最核心的問題可能是產品推廣不夠全面，造成客戶鮮少接收到關於這兩項產品的資訊。因此針對這兩項產品的行銷方式，目前首要的工作應當是要廣為宣傳，由於資料顯示幾乎所有客戶都會使用 G 通路，因此可以考慮使用 G 通路作為宣傳的管道。