2019 商業模式與大數據分析競賽

預測消費者信貸與基金之購買行為

GROUP22 競賽企劃書

指導單位：教育部
主辦單位：台新銀行
國立中山大學財務管理學系
國立中山大學管理學術研究中心
智慧電子商務研究中心
協辦單位：國立中山大學資訊管理學系

競賽摘要

競賽主題	**預測消費者信貸與基金之購買行為**
競賽摘要	近年因技術快速發展，全球數據生成與儲存量以高速成長，數據科學為商業營運新闢戰場。金融產業受到量化科技應用影響程度最劇的功能為技術研發與行銷管理。而掌握金融行為並利用技術將其轉換為商機、競爭優勢，為本團隊利用人力智慧進行創新商業模式之核心目標。行銷科技的一個面向為從行為面的數據捕捉到潛在顧客之需求，潛在顧客管理全面由能力導向走向技術引導。本團隊之人工智慧模型的應用不只在於提升決策效率或執行差異化策略，更重要的是基於客戶導向，提供即時商機達到極大化行銷效益與優化潛在客戶管理與效益。 本團隊使用台新銀行提供之客戶狀態、購買資料，利用近年來先進開源機器學習模型 CatBoost 來預測客戶購買行為，期以將消費者劃分為較可能購買與不太可能購買之類別，實行精準行銷，優化廣告投放，提升顧客命中率。 經過大量調整參數，訓練出來的模型，根據台新銀行給的資料，本模型於預測 2018 年 12 月之購買行為時 y1 達到了 F1_Score 為 0.1312，預測 y2 達到了 F1_Score 為 0.2985。最後把前五個月資料餵入機器訓練之後，再根據 2018 年 12 月之資料，預測 2019 年 1 月消費者之購買行為。

目錄

壹. 緒論

1. 研究背景與研究問題

根據投信投顧公會截至 2019 年 9 月底止，國內共計有 39 家投信公司，所發行的共同基金總數為 955 支，國內投信公司整體基金規模約為 3.7 兆元，較前月(2019 年 8 月)底規模 3.6 兆元，成長幅度 1.25%。海外基金截至 2019 年 9 月底止，共核准 41 家總代理人、67 家境外基金機構、1,032 檔境外基金，國內投資人持有金額約 3.6 兆元。可見台灣投資人對於基金購買力與投資意願高，且對一般投資大眾而言，相較於其他金融資產，共同基金投資的門檻比較低，訊息亦較公開。

金融商品業務係基於專業理財經理人的對於客戶質化敘述資料(投資標的偏好、風險趨避程度)的理解，輔以客戶資產水位，推薦個人化資產配置方案。台新投信目前基金商品(包括 ETF)數量為 53 支，規模約 920 億，產品線完整，滿足各類型風險偏好投資者的理財需求。 在高績效表現的同時我們也開始思考如何近一步提高金融商品投資服務的附加價值，達成「較高之單位勞動報酬」、「較高之營利表現」之特徵。我們提出由客戶主觀表達之投資偏好是否有過度自信或基於不完整資訊而導致之偏誤？在客戶存有決策與誤判可能性的條件下，該如何回應「較高之單位勞動報酬」、「較高之營利表現」目標。

根據經濟部「2019／2020 產業技術白皮書」之定義，高單位勞動報酬主要來源為企業具備高專業性或高技術性服務門檻，或應用先進科技輔助提高單位勞動生產效率。而高營利表現之關鍵來源，亦為藉由多元科技應用營造持久性競爭優勢。

2. 數據驅動型行銷模式

在金融產業競爭範疇中，產業參與者無不掌握巨量客戶行為，然而握有數據與精準行銷是行銷科技(MarTech)光譜的兩端，大多數人集中於數據資料端，少數產業領導者則在光譜的中間位置。金融商品之行銷廣告與顧客關係管理益趨精密複雜，潛在顧客管理日漸被重視，如何捕捉到在將來某一時點轉變為既有顧客的人，這是企業開拓市場、競爭市佔的關鍵力量。

根據商發院創模所之整理，金融服務範疇受到量化科技應用影響程度最劇的功能為技術研發與行銷管理。而整握金融行為並利用技術將其轉換為商機、競爭優勢，為本團隊利用人力智慧進行創新商業模式之核心目標。行銷科技的一個面向為從行為面的數據捕捉到潛在顧客之需求，潛在顧客管理全面由能力驅動(Capability-Driven)走向技術引導(Technology-Led)，結合奠基於實際資料分析與行銷科技工具應用。本團隊之人工智慧模型的應用不只在於提升決策效率或執行差異化策略，更重要的是基於客戶導向，提供即時商機達到極大化行銷效益與優化潛在客戶管理與效益。

貳. 探索性資料分析

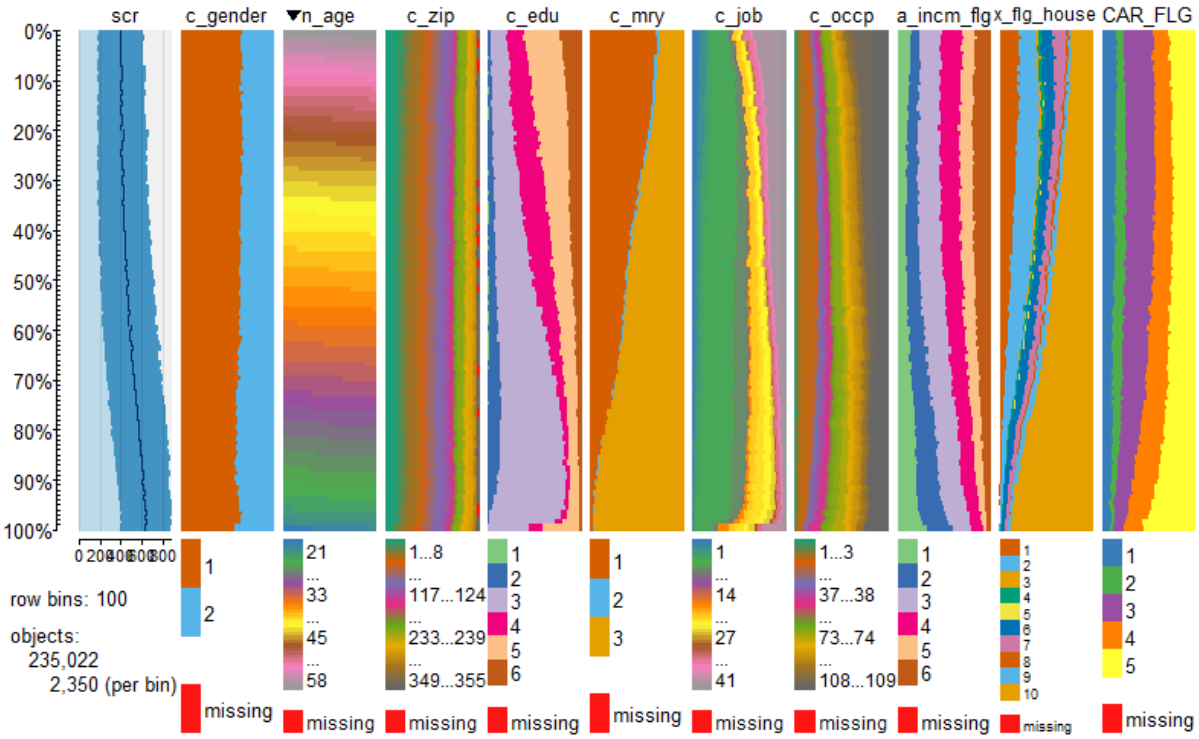
1. 讀取資料

共有五個資料表，分別是 profile.csv, status.csv, sr_1.csv, result_y1.csv,result_y2.csv。profile 共有 235,022 筆資料，status 有 1,410,132 筆，剛好是 profile 的六倍。每個消費者有六個月的資料。

觀看 y1 與 y2 資料表，y1 只有 4,786 筆購買紀錄，y2 只有 3,295 筆，為非常類別不平衡之資料，即有購買跟沒購買的人的比例非常懸殊。

2. 概覽各變數資料分布情況

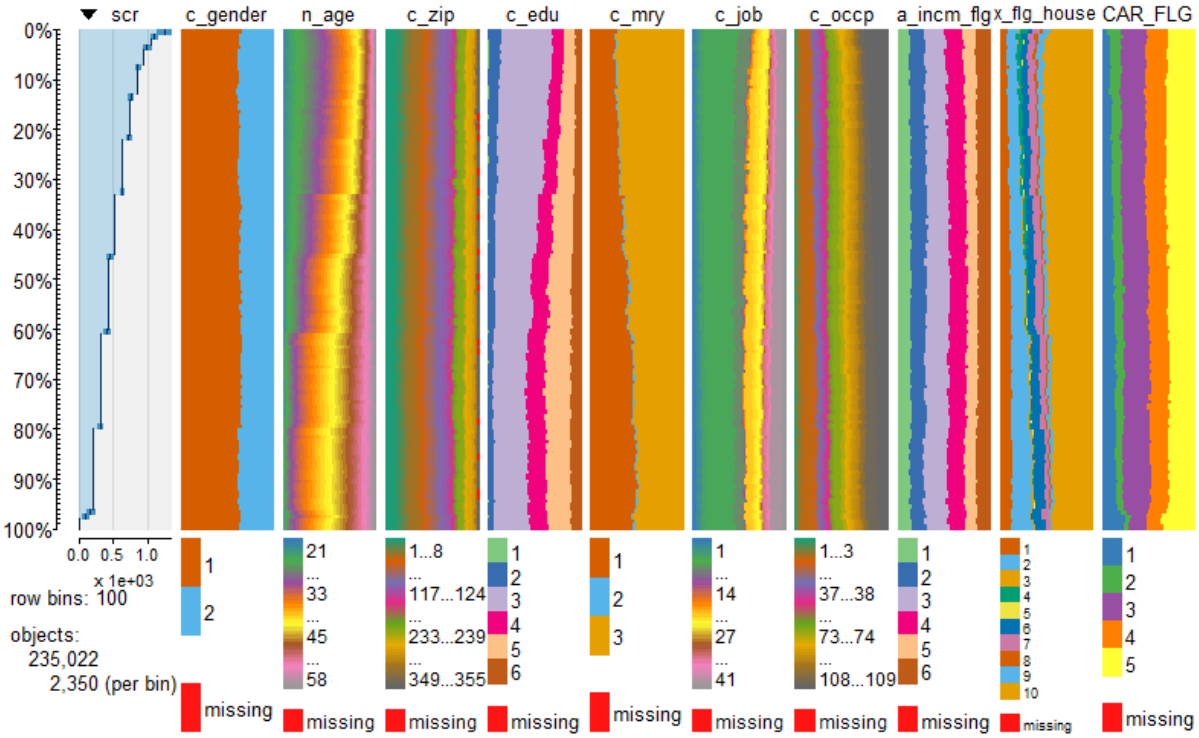
使用 R 當中的 tableplot function，我們首先依照年齡大小對應其他各個顧客背景變數製圖，去推估所給資料的類別屬性，結果如下：



我們可以大致得到以下推論：

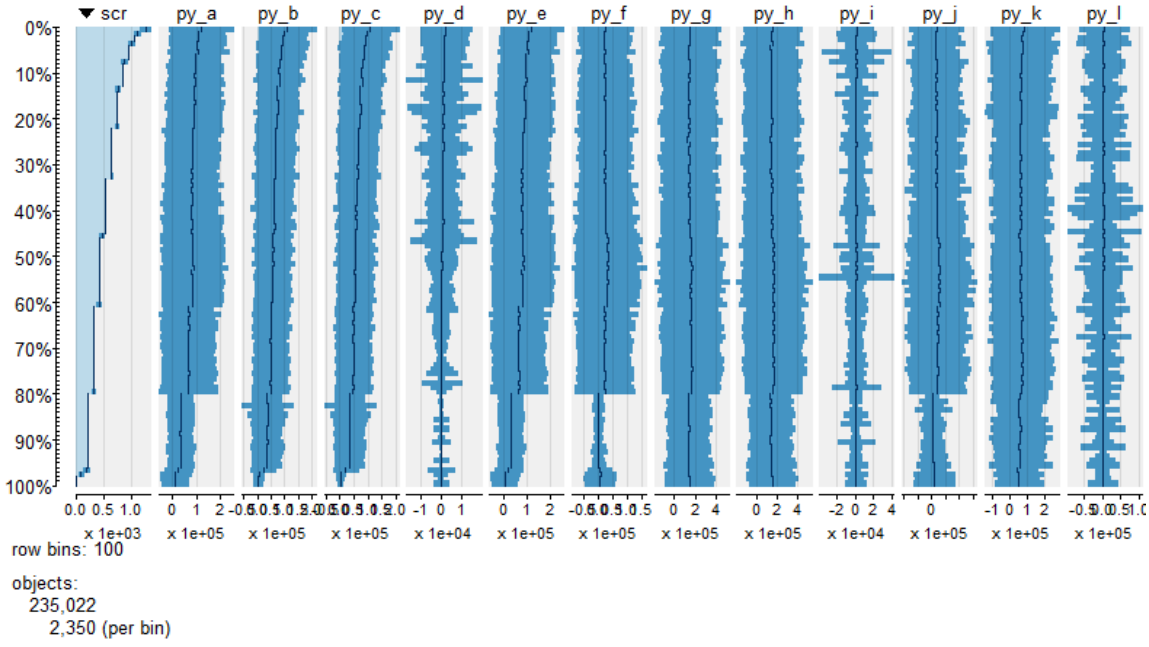
c_edu 越小表示教育程度低，c_mry 為 1 表示已婚、3 表示未婚，a_incm_flg 同資料所給描述，越高表收入越高，x_flg_house(不動產狀態代碼)和 CAR_FLG(動產狀態代碼)我們則推估是數字越小表此項資產越高。

進而我們以忠誠度 scr 分數的高低進一步檢視其他客戶背景資料的分布情況，如下：



可以發現 scr 分數在不同性別、居住地區、工作屬性、職位類別、年收入等級、動產狀態類別上並沒有明顯的差異，但在年齡、教育程度、婚姻狀況、不動產狀態類別上有不同。我們大致可以看到越年輕、教育程度較低、未婚、不動產狀態類別小的族群在 scr 的分數越高。

另外我們也拿 scr 分數的排序高低去看不同信用卡消費類別的分布狀況，製成圖如下；



我們可以大致看到 scr 的分數越高，多數的 py 類變數也越高，除了 py_i 和 py_l 這兩類變數沒有類似的分布情況。其中又以 py_a、py_d、py_e、py_f 凸顯此類狀況，是值得我們細部去看的。

3. 辨認遺失值

發現 c_zip 這個變數有遺失值，這邊直接把所有遺失值轉換成 0 處理，代表該人沒有 zip。

4. 特殊資料判斷處理

(1) Profile

多數為類別資料，其中可以看到有數個 Categorical Variable 像是 c_zip 有三百多種的值，可見需要特殊處理(使用 CatBoost 模型即解決)。

(2) Status

自有資產類別資料，as_X，因為是該使用者在該時間的狀態，是為一個存量，與其他類別變數不同，常有在六個月的資產固定不變的值，故也應特殊處理(後面處理為資產的變化量 as_X_flow)。

(3) Sr_1.csv

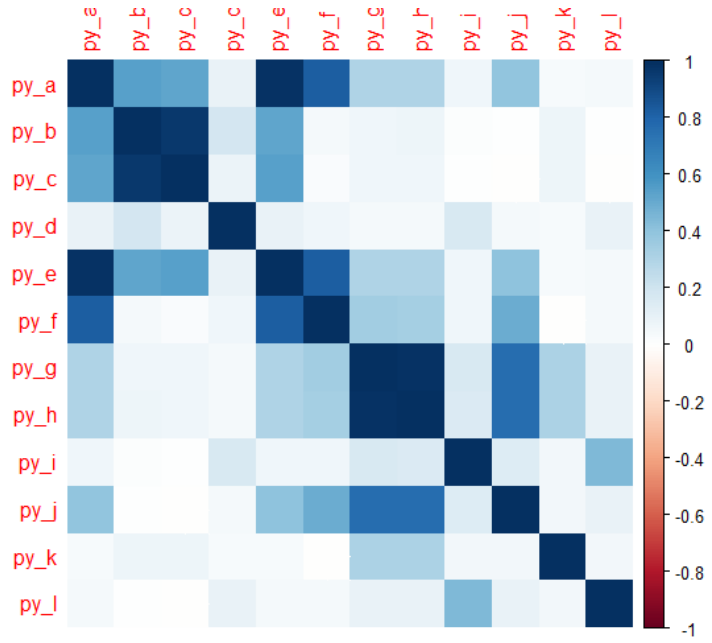
其每人每月之消費總結，都總結成 status 資料表了，加上因消費品項這種文字型的資料需特殊處理(Natural Language Processing,NLP)，因此本組後面模型並沒有使用到此資料表。

5. 辨認異常值(過於違反常理的值)

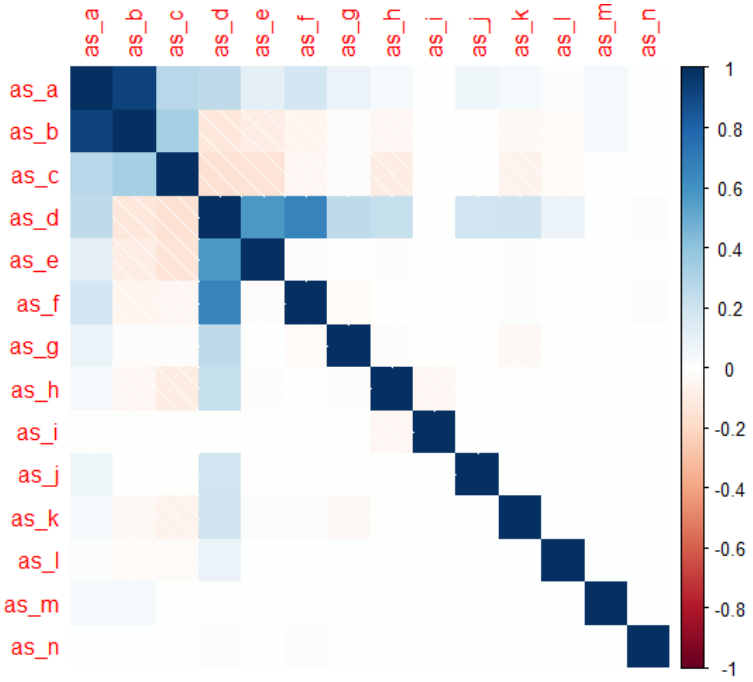
藉由查看分佈與敘述統計的方法，沒有看到任何須處理之異常值。

6. 變數間相關性

我們將 py 類的變數兩兩看相關性，探測是否有任何變數彼此之間相關性過高，亦即有高度共線性，使得後面模型處理上產生問題，呈現如下：



我們發現 py_a 和 py_e 彼此之間相關性異常的高，回頭去看原始資料，發現 py_a 實為 py_e 往後挪一個月的資料。同時 py_b 和 py_c 之間也具有高相關性，回頭看原始資料，發現雖不若 py_a 和 py_e 之間為前後期重合的資料，但 py_c 也與下一期的 py_b 相當近似。而 py_g 和 py_h 如同前面的 py_a 和 py_e 為前後期關係，py_h 實為 py_g 往後挪一個月的資料。



我們另外將 as 變數兩兩去看他們的相關性，此處 as 變數透過將第 12 個月(期末)減去七月份資料(期初)轉換為變化量的概念。我們發現 as_a 和 as_b 幾乎相同，除了少數的情況 as_b 會比 as_a 來的小。另外 as_e 和 as_f 的變化量加總會為 as_d 的變化量。

參. 特徵資料處理過程(詳細請點此連結，Github)

1. 資料前處理

(1) 針對二元類別變數進行轉換

將 profile.csv 之 c_gender, c_mry 轉為 0,1 值。

(2) 計算每人每月信用卡消費總額與資產持有總額

將 status.csv 之 py_X 變數與 as_X 變數，每個月份總合起來，命名為 py_total 與 as_total 特徵。即代表每個月份消費與資產持有的總和。

(3) 將 result_y1.csv, result_y2.csv, profile.csv, status.csv 融合成 merge.csv

根據 resulty1 與 resulty2 做處理，有購買的月份設 1，沒購買的月份設 0。最後把這四個資料檔，根據"srno"這個變數做融合(outer join)。

	srno	YYYYMM	y2	y1	py_a	py_b	py_c	py_d	py_e	py_f	py_g	py_h	py_i	py_j	py_k	py_l	scr	as_total	c_gender	n_age
0	53385	201807	0.0	0.0	0	0	0	0	0	0	22154	25935	0	0	0	0	210	253780	1	48
1	53385	201808	0.0	0.0	0	0	0	0	0	0	7694	22154	0	0	0	0	210	253780	1	48

(4) 特殊處理

本組在融合 Merge 的時候，把 Y 值處理成該月有購買的指標，因此如果這樣直接餵進模型學習，模型學出來的意義是：

根據該月消費者的狀態，預測該月消費者會不會購買

然而，本比賽是要預測 2019 年 1 月的購買行為，但不會有 2019 年 1 月的消費者資料，只有 2018 年 7 月至 2018 年 12 月的資料。

因此本組將所有 y 值往前移一個月份，最後一個月(201812)設定為 0，這樣的做法實際上便是把意義轉換為：

根據該月消費者的狀態，預測下個月消費者會不會購買

2. 特徵工程

(1) 刪除 as_i 與 as_h

這兩個變數為非常稀疏之矩陣，充滿了大量的零與極少數的值。以這種方式訓練模型只是造成機器的負擔，加上該少數的值與 y1 也沒什麼特別的關係，因此直接剔除。

(2) 刪除 YYYYMM，時間資料

因為後面使用到之 CatBoost 模型並不是一個時間序列模型，不會用到時間資料，因此予以刪除。

(3) 計算每個月份的各類資產與消費流量變化

定義：流量變化為這個月份的量減掉上個月份的量
算出兩種 flow：py_X_flow 與 as_X_flow。
因此，第一個月的流量變化為 0(沒有 2018 年 6 月的資料，無法得知 2018 年 7 月之流量)。根據該定義，算出每個人每個月消費與資產流量變化。

肆. 預測模型建構與方法

1. 演算法選擇

由於數值型(numerical)資料較好處理與使用模型，因此不特別討論此類別資料。根據此資料的類型與 EDA 後的結果，發現不少重要之變數為類別資料，因此如何處理類別變數便是模型選擇的重點。
經過大量研究與測試後，選擇了處理類別變數性能較好的模型 CatBoost。接下來以簡潔簡短的方式介紹 CatBoost。

- CatBoost 簡介

CatBoost 基於 Gradient Boosted Decision Trees，是 Yandex 開發的一種新的機器學習技術，其性能優於許多現有的增強算法，如 XGBoost、Light GBM 等。

- CatBoost 原理(原論文連結) – Target/Mean Encoding

即使用對具有相同分類特徵的所有數據點的目標值的均值來表示每個分類特徵，Catboost 額外根據每個資料點加入時間以處理 Target Encoding 之目標洩漏(Target Leakage)問題。

2. 建構模型過程

將處理且合併好之 merge.csv 讀取

(1) Train/Validation/Test split

使用前四個月的資料當 Train，第五個月資料當作 Validation 下去訓練模型，最後一個月當作 Test。

處理 Class Imbalance 問題：在 Train 與 Validation 抽樣的時候，根據 y 之比例做分層抽樣。即：Train 與 Validation 之 y 比例會一樣。

(2) 訓練模型

辨認所有類別變數，以利 CatBoost 得知哪些變數應做類別變數的處理，Loss function 使用 Log loss，以 F1_Score 做衡量，且把 Positive Sample(有買 y 的資料點)之權重設為 100 做訓練(一樣為了處理 Class Imbalance)。

(3) 調整參數

原本有調整一些其他的參數(如 regularization coefficient、depth 等)，但因為訓練出來都使 F1_Score 下降，故最後還是使用 CatBoost 預設參數。

```
params = {
    'iterations': 200,
    'learning_rate': 0.1,
    'loss_function':'Logloss', #CrossEntropy?
    'eval_metric': 'F1',
    'random_seed': 1337,
    'logging_level': 'Silent',
    'class_weights' : [1, 100 ],
    # 'depth' : 10 調過了，感覺變爛
    # 'l2_leaf_reg' : 5 增加=>worse
    # one_hot_max_size  調過了，感覺沒有什麼幫助
    #'scale_pos_weight' : 10   #(1410132-4786)/4786
    #,'use_best_model' : True
}
```

伍. 模型準確度與結果分析

根據以上預測模型之建構與 Train/Validation/Test split 之資料，訓練出來之模型來預測第五個月的 Y 值(即第六個月會不會購買)。

Y1 之 F1_Score 結果如下圖(Threshold 設 0.62)：

```
In [180]: threshold = 0.62 # threshold we set where the probability prediction must be above this to be classified as a '1'
          classes = predictions_probs.copy()[:,1] # say it is the class in the second column you care about predictint
          classes[classes>=threshold] = 1
          classes[classes<threshold] = 0

In [183]: f1_score(y_test, classes)

Out[183]: 0.13122529644268777
```

Y2 之 F1_Score 結果如下圖(Threshold 設 0.9)：

```
In [171]: threshold = 0.9 # threshold we set where the probability prediction must be above this to be classified as a '1'
          classes_y2 = predictions_probs_y2.copy()[:,1] # say it is the class in the second column you care about predictint
          classes_y2[classes_y2>=threshold] = 1
          classes_y2[classes_y2<threshold] = 0

In [174]: f1_score(y_test, classes_y2)

Out[174]: 0.2984913793103448
```

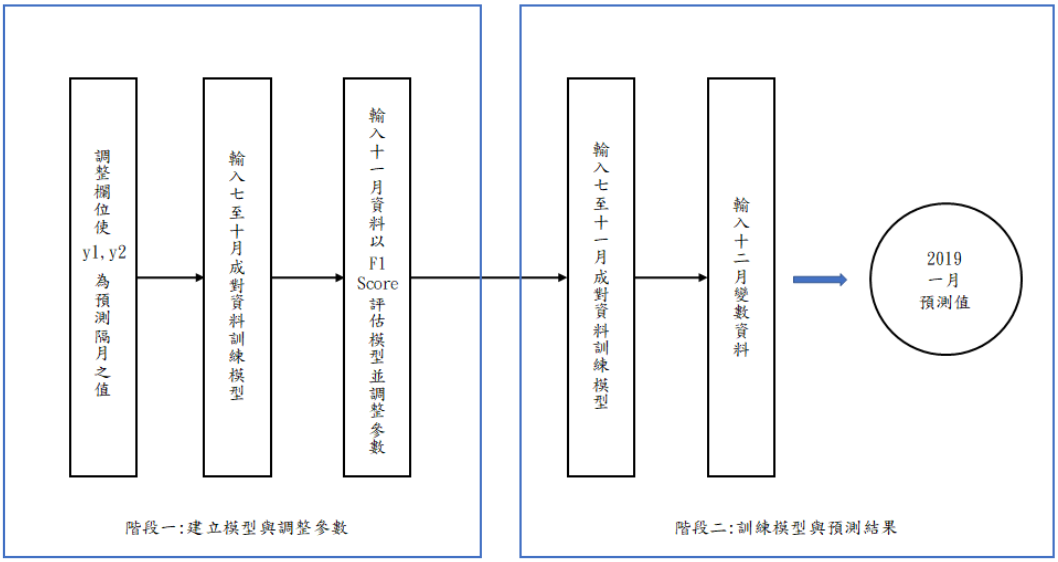
(更多詳細細節請參照 GitHub)

陸. 生成模型預測 2019 年 1 月之購買行為

Train/Validation/Test split :

這邊改成訓練前五個月的資料，預測第六個月的 y 值，因此
Train/Validation set = 前五個月資料，Test set = 第六個月的資料
因為沒有第六個月的 y 值(如上所述， y 值為該人「下一個月」會不會購買)，即 2019 年 1 月之購買資料，因此無法計算 F1_Score。

Y1、y2 分別做預測，將預測之結果根據"srno"合併成結果 submission.csv。



建模流程總結