封面

第二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

隊伍組別：第 17 組

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

中 華 民 國 108 年 10 月 25 日

競賽摘要

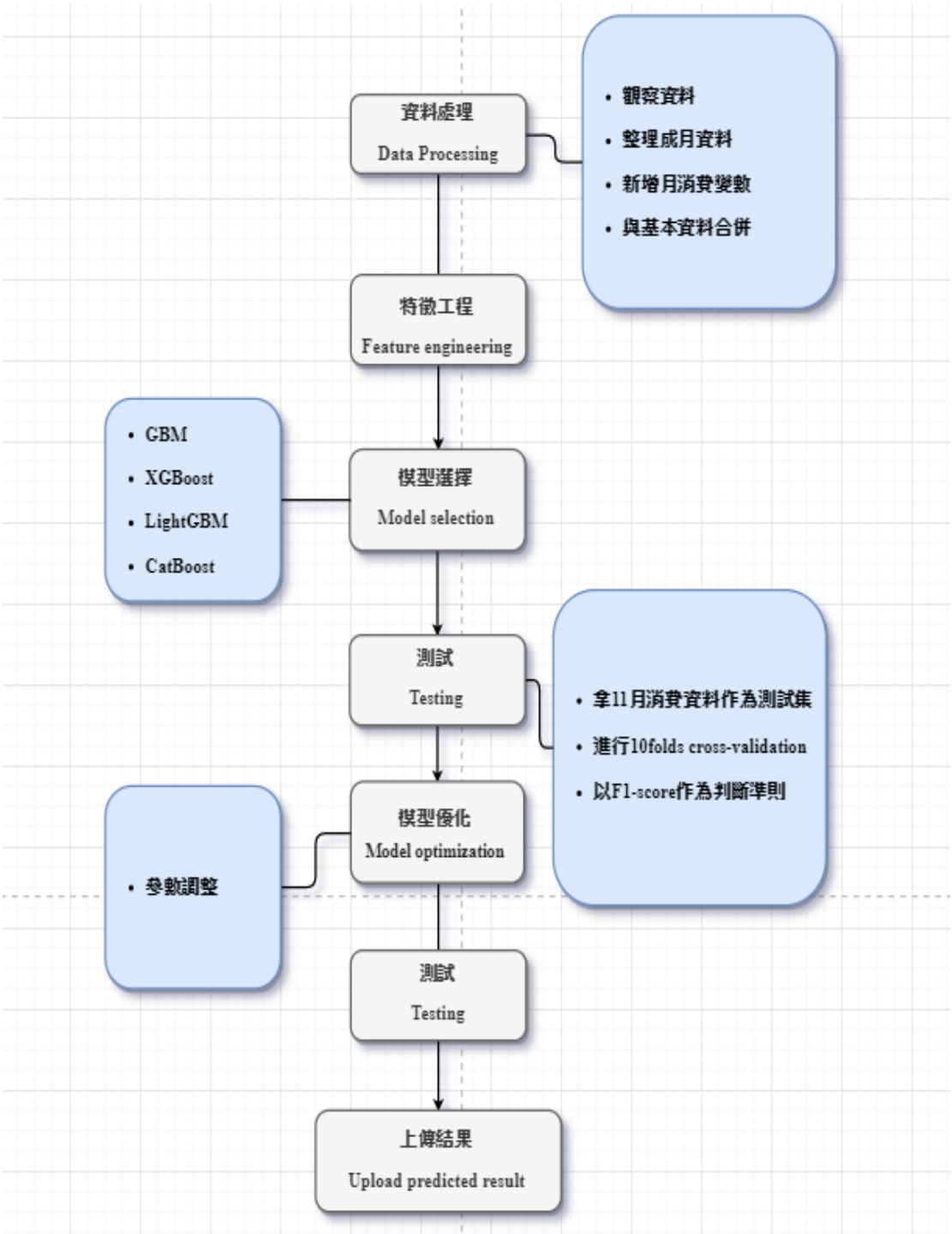
競賽主題	最適金融商品預測
競賽摘要	根據我們模型預測的理念，我們以前一個月的資料去預測下個月是否會購買這項產品。在此我們用了7-10月的資料及8-11月的Y當作訓練集，11月的資料及12月的Y則作為測試集。 在特徵工程上，我們將srl中每人的的消費金額，以月為單位取平均:每人每月出現最多次的mcc;觀察連續型變數後，將連續型變數彼此相關性較高的作主成分分析，提取有解釋能力的主成分做為變數...等 我們使用我們整理好的資料去跑XGboost、Naïve Bayes、GBM、Light GBM、CatBoost這五個模型，以F1-Score當作我們的評測標準。 我們希望能運用我們的模型，除了預測到會購買的人真的有購買之外，還希望能找出下個月的潛在客戶，也就是預測他們會購買，但事實上沒有購買的客戶群。針對這群潛在的客戶群，以「了解你、認識你、擁抱你」為理念提供更適合的行銷方式。除此之外，在預測完客戶是否會購買的預測之後，針對預測出會購買的潛在客群進行分群，抓取出他們的特徵，以利於我們針對各客群做精準行銷；另外，除了針對潛在客戶群，我們也希望能更了解不會購買的人，依照「三你」的理念，充分理解客戶不購買的理由，加以改進產品與發展新的行銷方式，以客戶為本，希望達到以客戶推廣客戶的方式，慢慢擴大企業在社會上的客戶信任度。

人工智慧金融挑戰賽隊伍 1

分析說明書

目錄:

一、流程圖：


資料處理
Data Processing
觀察資料
整理成月資料
新增月消費變數
與基本資料合併
特徵工程
Feature engineering
GBM
XGBoost
LightGBM
CatBoost
模型選擇
Model selection
測試
Testing
拿11月消費資料作為測試集
進行10folds cross-validation
以F1-score作為判斷準則
模型優化
Model optimization
參數調整
測試
Testing
上傳結果
Upload predicted result

（圖一 流程圖）

二、資料處理：

1.

考慮到我們已經有的標籤是 8 至 12 月是否有回應 y1、y2 產品，因此第一部份我們就先將所有資料整理成月份資料：

（1）將 sr1 中，每人的的消費金額，以月為單位取平均

（2）每人每月出現最多次的 mcc

（3）每人每月是否有信用卡消費、信用卡消費平均、最大值

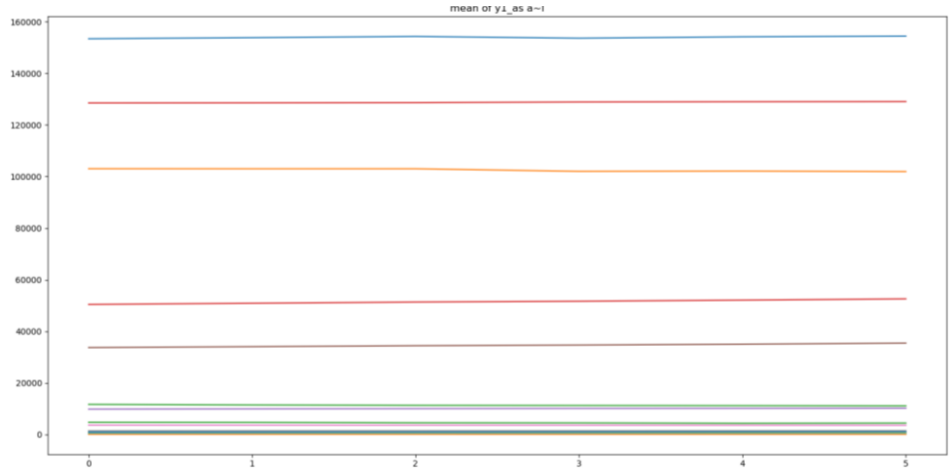
（4）每人每月是否持有資產，持有資產金額平均、最大值

（5）以客戶編號及月份為依據，將上述四項與 status 合併，一列就是單一客戶在某個月的消費資料。

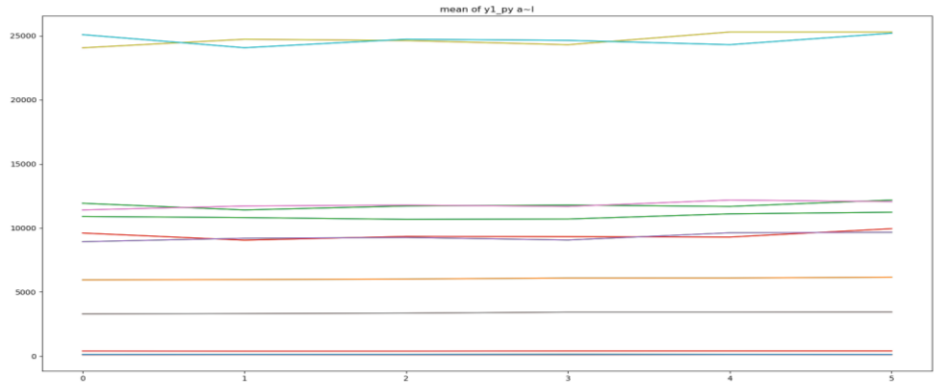
2.

Profile 是不會隨月份變動的基本資料，我們把客戶基本資料與每月消費相關資料合併成一欄，也就是單一客戶的客戶資料會在七月那一列、八月那一列…都出現一次。

本次比賽的目的是預測一月是否會購買，我們想建立一個以前一月的資料預測後一個月是否會購買的模型。



（圖二 As 系列連續型變數平均對月數）



（圖三 PY 系列連續型變數平均對時間）

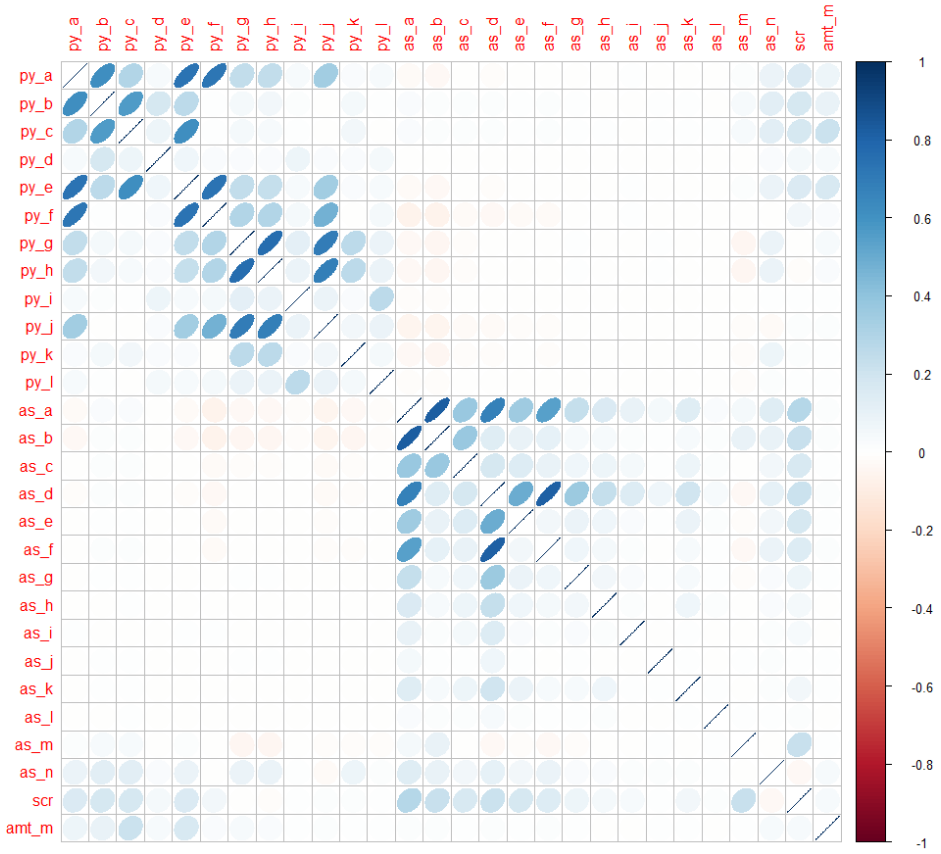
如上圖，針對連續型變數，我們觀察每個月所有客戶的平均，並沒有發現隨著月份變化的趨勢。

最終決定把每位客人在每個月的消費相關資料視為獨立的樣本，客戶每個月的消費資料對應隔月的 y1、y2 產品回應紀錄，進行模型訓練，最後我們使用 7~11 月的消費紀錄，得到一個總共有 1175110 個樣本的資料集，並以 12 月的客戶消費資料預測 1 月是否會回應。

三、特徵工程：

模型中的解釋變量之間由於存在精確相關關係或高度相關。看似相互獨立的變數本質上是相同的，是可以相互代替的，一般可能有一定程度上的共線性。

考慮到這點，我們將連續型變數提出，繪出相關係數矩陣熱圖(cor plot) (圖四)後，發現有許多連續型變數彼此相關性過高，可能導致某單一變數解釋力降低。



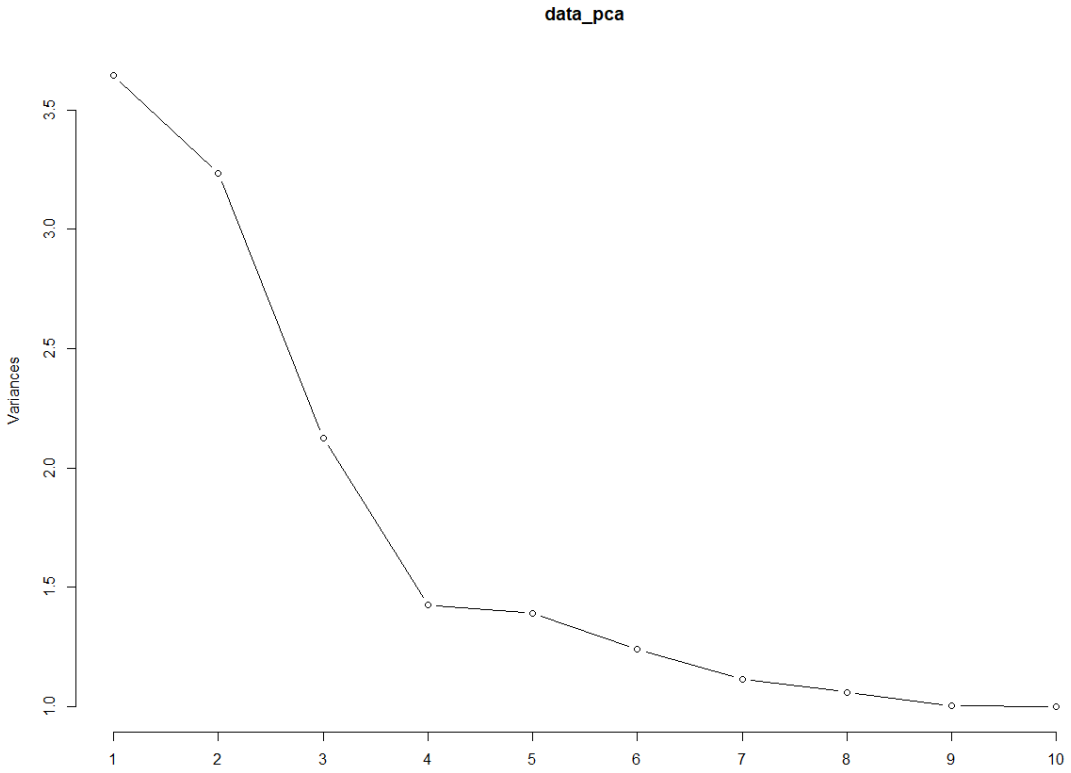
(圖四 相關係數矩陣熱圖)

主成分分析(PCA)經常用於減少數據集的維數，同時保持數據集中的對變異數貢獻最大的特徵。通過保留低階主成分，忽略高階主成分做到的。這樣低階成分能夠保留住數據的最重要方面。

而當變數間有較強的線性相關時，利用主成分分析法, 可以消除各變數之間的共線性或高度相關性, 減少變數的個數,利於我們後續分析。

因此我們嘗試將連續型變數做主成分分析，挑選出適合的主成分，以解決

變數彼此相關性過高的問題，而觀察陡坡圖(圖五)後，我們選取 4 個主成分與原始資料中的類別型變數做結合最後加上 y1 與 y2 結果，整合成新的資料級提供模型做嘗試。



(圖五 主成分分析陡坡圖)

	PC1	PC2	PC3	PC4
py_a	0.416926051	-0.069408755	0.234395006	-0.064243526
py_b	0.210894353	-0.094799679	0.405854896	-0.029578028
py_c	0.212751578	-0.098092629	0.417113075	-0.027387482
py_d	0.056064304	-0.015010371	0.090654373	-0.004999830
py_e	0.418260270	-0.069816848	0.236542836	-0.062286658
py_f	0.382422265	-0.009908991	-0.021989893	-0.061970301
py_g	0.341349413	0.001381837	-0.382688366	0.088624299
py_h	0.337898432	0.001151624	-0.381104535	0.087322802
py_i	0.066331847	0.002071238	-0.089491307	0.043273817
py_j	0.364753758	0.008840261	-0.342535783	0.067644631
py_k	0.098106632	0.011704559	-0.143933537	0.047010903
py_l	0.062421954	0.002879971	-0.085222880	0.036668061
as_a	-0.077818548	-0.506161768	-0.068876183	0.184477678
as_b	-0.063956620	-0.321236341	0.005875332	0.557356677
as_c	-0.033110312	-0.245178888	-0.017668512	0.372751350
as_d	-0.052735654	-0.465875838	-0.127686669	-0.401042439

(圖六　主成分與原變數的線性組合(部分截圖))

四、模型選擇和模型優化

根據我們模型預測的理念，我們以前一個月的資料去預測下個月是否會購買這項產品。在此我們用了 7-10 月的資料及 8-11 月的 Y 當作訓練集，11 月的資料及 12 月的 Y 則作為測試集。

我們使用我們整理好的資料去跑 XGboost、Naïve Bayes、GBM、Light GBM、CatBoost 這五個模型:

XGboost：

首先運用 10 次交叉驗證去尋找較佳的參數，但 test 跑出來的 F1-score 並不是很滿意。因此，我們認為是不平衡的資料所影響的，所以進行 undersampling 得到 1:1，1:2，1:3 的三組樣本，同時交叉驗證確定參數後，利用此參數組合建立模型，但 test 跑出來的 F1-score 並不是很滿意。

GBM：

我們加入新的特徵工程有當月有購買 y2 的變數，將 gender 和 marry 做排列組合成新的變數，py_a-l 做加總做為新變數。除此之外，因為變數重要性中 amt_m 排名第三，所以將 amt_m 放大十倍權重後然後新增這個變數嘗試。發現 smo 客戶編號變數重要性也很高，故不將此排除。在測試 Y2 時，新的特徵工程有當月有購買 Y1 的變數，其餘皆與測試 Y1 時相同。在跑模型時，我們並沒有做 undersampling，我們運用 10 次交叉驗證去尋找較佳的參數，但 test 跑出來的 F1-score 並不是很滿意。

Naïve Bayes：

這次分為兩個面向去進行。首先是純粹對原始資料將 7-10 月當作 train，11 月當作 test，但 test 跑出來的 F1-score 並不是很滿意。在我們試過多個模型之後，F1-score 都沒有顯著的改善，因此我們懷疑是有一些 noise 存在，因此我們將資料中的連續變數做 PCA 後合併類別變數，但 test 跑出來的 F1-score 仍然不是很滿意。

LightGBM：

將 7-10 月當作 train，11 月當作 test，在測試 y1 的時候，將 y1 = 0 的資料作重抽樣，抽到資料筆數分別是 y1 = 1 的二十倍，再將重抽樣的資料分別與 y1 = 1 的所有資料合併作為訓練資料。測試 y2 的時候也是相同的方法。我們運用交叉驗證去尋找較佳的參數，確定參數後，利用此參數組合建立模型，但 test 跑出來的 F1-score 並不是很滿意。

CatBoost：

我們經歷過了多方測試及參數校正，加上模型變數重要性之篩選以及決策點的測試及運算，得到一個測試結果最為良好的模型當作我們的最終參數設定。其中最為重要的參數調整便是「權重」的設定了，因為 Y 本身呈現一個極度不平衡的趨勢，透過權重的設定來凸顯 Y 的特徵，最終才能跑出一個有良好預測結果的模型。也是我們的最終模型選擇

五、結論：

最後我們選擇 Catboost 模型來當作最終模型，並固定我們調整好的參數將 7-11 月的資料來去做訓練，並以 12 月的資料來預測產生 Y1、Y2。

在我們的研究過程中我們也發現，Y1 的 F1-Score 普遍偏低，變數重要性也並未特別突出，因此合理懷疑 Y1 在此資料中並沒有顯著特徵。

而在最後模型中，我們將客戶編號也加進去，因為他在單月上變數相對重要，又考慮到客戶編號有可能跟顧客成為會員的先後時間順序有關，所以將此變數也加入模型；若更進一步了解資料後，客戶編號單純為亂數流水號，則將此變數移除。