

巨量資料分析期末 Project #1

玉山人工智慧公開挑戰賽 2021 冬季賽

信用卡消費類別推薦

巴斯克乳酪蛋糕

統計 111
H24076029
劉米婷

統計 111
H24071215
陳柔漪

統計 111
H24071273
林少穎

Table of Contents

1. Problem Statement.....	2
1.1 競賽說明與目標.....	2
1.2 競賽評比指標.....	2
2. Data preprocessing.....	2
3. Insights discovered from the data.....	3
3.1 遺失值紀錄 (以全部資料做計算).....	3
3.2 消費次數與類別關係	3
3.3 消費類別與年齡.....	4
3.4 消費類別與性別.....	4
4. The method you designed	5
4.1 僅使用總金額進行計算 (不使用模型).....	5
4.2 目標變數為 TOP3 消費類別	5
4.3 目標變數為 TOP1 消費類別	6
4.4 計算 16 種消費類別的消費總金額與總次數	6
4.5 計算 16 種消費類別的消費總金額與總次數 + 加入其他變數	7
5. Results	7

1. Problem Statement

1.1 競賽說明與目標

使用兩年內的顧客基本資料與信用卡消費紀錄，去預測每位顧客下個月份消費金額前三名的消費類別排序。原始消費類別總共有需要預測的類別指定為其中的 16 種類別。

1.2 競賽評比指標

此次競賽採用 $NDCG@3$ (Normalized Discounted Cumulative Gain) 作為評分標準， $NDCG$ 的計算如以下。

$$NDCG = \frac{\sum_{c \in C} NDCG_c}{|C|} = \frac{\sum_{c \in C} \frac{DCG_c}{iDCG_c}}{|C|}$$

C 為所有需要預測的顧客，若 c 為其中一位顧客，則 $NDCG_c$ 為顧客 C 之 Normalized Discounted cumulative gain， DCG_c 則為顧客 c 之 Discounted cumulative gain，而 $iDCG_c$ 為顧客 c 最理想之 Discounted cumulative gain。

由此評分可知需預測的前三名消費類別中，第一名的預測準確度佔了大部分的重要性。

2. Data preprocessing

資料集可分為使用者「信用卡消費資料」與「基本屬性資料」，全部有 53 個特徵變數，資料總筆數為 32975653 筆，而顧客數有 500000 人。

因為資料量非常龐大，我們首先決定僅保留需預測的 16 種消費類別資料往下探討。再者，資料的總時長有兩年，我們覺得顧客喜好會隨著時間有所改變，因此時間上我們也只保留了第二年 (dt 13 ~ 24) 的消費紀錄，而越靠近預測月的時間是相對重要的。

要預測的變數為消費類別，且一位消費者可能具有多種類別的紀錄，為了讓預測更貼近前三名，因此我們對於消費類別資料有做新的整理，大致分成以下種類，詳細資料格式置於後方第 4 部分。

- 取消費金額前三名的類別
- 取消費金額第一名的類別
- 以消費類別為單位加總該類別的金額與次數

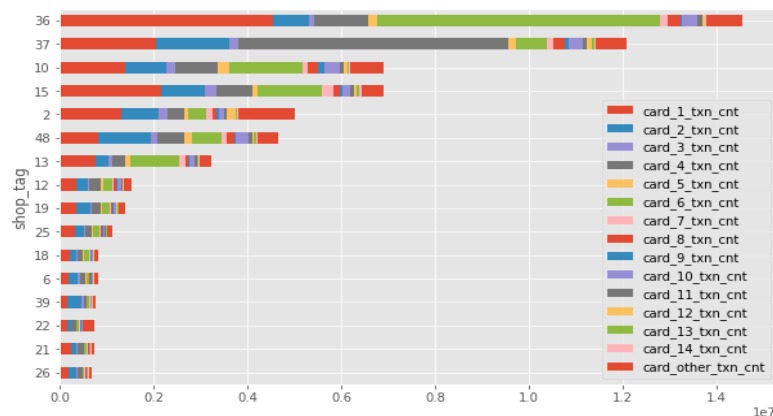
3. Insights discovered from the data

3.1 遺失值紀錄 (以全部資料做計算)

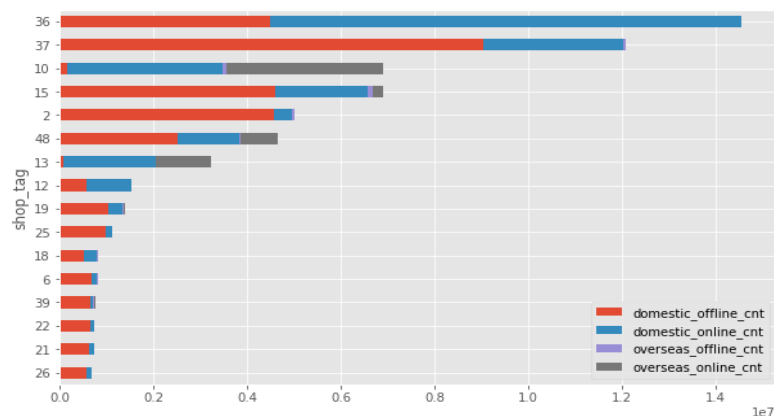
變數	筆數	變數	筆數
masts	10 (<0.01%)	cuorg	10 (<0.01%)
educd	23 (<0.01%)	slam	1749906 (5.3 %)
trdtp	435 (<0.01%)	gender_code	421667 (1.3 %)
naty	10(<0.01%)	age	421667 (1.3 %)
poscd	10 (<0.01%)		

3.2 消費次數與類別關係

信用卡資料記錄 14 張卡的消費次數，以及國內外、線上線下購物次數，將這些變數與消費類別進行繪圖。可以看到類別 36 與 37 是擁有最多的消費次數，圖一中顯示類別 37 在 4 號信用卡有相對多的消費紀錄，圖二顯示購買類別 10 與 13 的商品較不容易從國內線下交易取得。



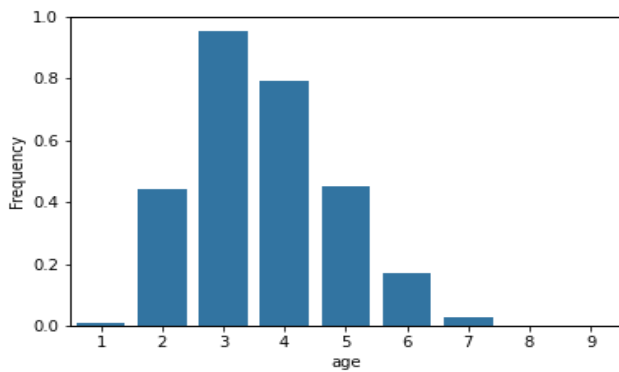
▲圖一，消費類別與 14 張信用卡的消費次數



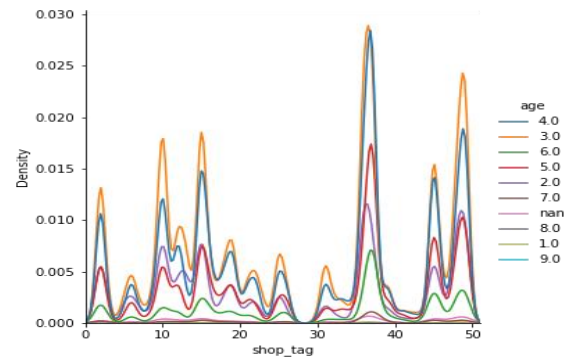
▲圖二，消費類別與國內外、線上線下購物次數

3.3 消費類別與年齡

我們認為不同的年齡層會擁有不同消費請傾向，故將年齡與消費類別繪製分布圖，可以看到此份資料主要為年輕人居多，第四區年齡在 37 類別上也有較明顯的升高。



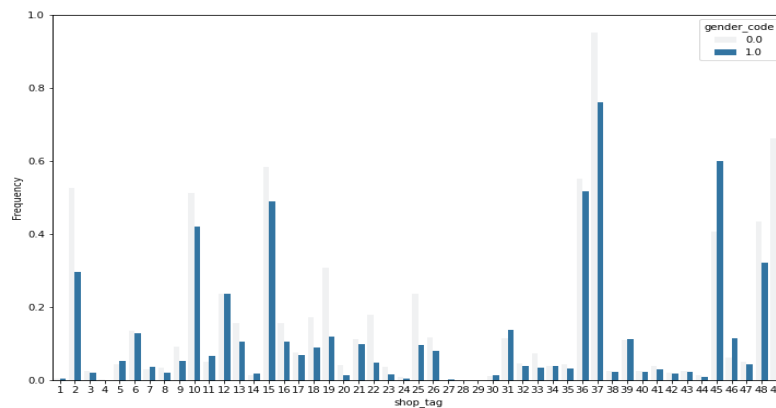
▲圖三，年齡分布



▲圖四，年齡與消費類別的分佈關係

3.4 消費類別與性別

接著來看性別與消費類別的關係，在指定的 16 個消費類別中，6、12、39 類別有相對較平衡的性別比例，其他類別都是由女性購買多過男性。



▲圖五，性別與消費類別的關係

4. The method you designed

4.1 僅使用總金額進行計算 (不使用模型)

- 方法描述

Top3 消費類別的產生來自於消費總金額，因此我們嘗試取用最近的四個月，將每個顧客在每個消費類別的總金額做加總，並以此建立 top 3 消費類別，此段時間內無消費記錄的顧客使用 36, 37, 15 (消費類別的總 top 3) 去填補。

- 上傳結果

Public	Private
0.6375	0.6358

4.2 目標變數為 TOP3 消費類別

- 模型

LGBMClassifier + Multioutput Classifier(from sklearn)

- 方法描述

以一個 ID、三個月為一組產生特徵變數 X，特徵變數主要包含一個顧客的前三消費類別，不足三個類別以指定 16 個消費類別隨機補充，其餘消費紀錄以加總處理。則在三個月消費類別的情況下一共產生 136 個特徵欄位。

	dt	ID	tag 1st	tag 2nd	tag 3rd	Feature...	時間紀錄	tag 數量
X	13~15	1	1	2	NA	...	1	2
	16~18	1	1	NA	NA	...	2	1
	19~21	1	1	6	2	...	3	3

因為競賽目標需要預測出三個消費類別，因此我們將第四個月的 Top3 消費類別作為目標變數 y，並使用 sklearn 的 Multioutput Classifier 輔助我們進行三維資料的預測。而後我們為了提高預測的正確程度，在資料量的足夠的形況下，若該筆資料無法透過加總計算得到 y，我們就捨棄該筆訓練資料。

	dt	ID	tag 1st	tag 2nd	tag 3rd
y	16	1	37	36	2
	19	1	37	36	2
	22	1	37	2	36

- 上傳結果

與單純計算消費金額有點落差，研究後發現我們所選擇的 Multioutput Classifier，更多使用於多標籤估計，較不適用於 Ranker 類型的題目，故往下嘗試進行別種方法。

Public	Private
0.5200	0.5184

4.3 目標變數為 TOP1 消費類別

- 模型

LGBMClassifier + `predict_proba`

- 方法描述

資料建構與 4.2 相同，但分數上並不是很好，於是參考同學期中報告的做法將目標變數 y 改為使用 Top1 消費類別，並使用 LGBMClassifier 模型計算出顧客在 16 種類別的機率，取前三高類別做為答案。此方法因為顧客擁有的 Top1 消費類別較為完整，在目標變數上能保留更多的資料筆數。

- 上傳結果

改換模型的使用方法後，分數明顯增加，但依然不如使用消費金額下去估計的高。因此，我們決定改換資料格式，希望能更著重在單個顧客的消費金額與類別上的連結。

Public	Private
0.5798	0.5787

4.4 計算 16 種消費類別的消費總金額與總次數

- 模型

LGBMClassifier + `predict_proba`

- 方法描述

經由上方方法得知使用 LGBMClassifier 配合下個月的 Top1 消費類別是可行的，為了更注重每個消費類別的消費金額與次數，我們加總每位顧客每三個月內 16 個消費類別的總金額與總次數，做為特徵變數(如下圖)。時間切割上每三個月一組，之間重疊一個月，即 dt 13~15、15~17...，則測試集使用 dt 22~24。目標變數一樣為第四個月的消費類別 Top1，無資料的顧客約有 31000 人，使用出現頻率前三高的消費類別進行填補。

總消費次數								總消費金額		
	chid	2_cnt	6_cnt	10_cnt	12_cnt	13_cnt	...	19_amt	21_amt	22_amt
0	10000000	0.0	0.0	7.0	0.0	0.0	...	0.0	0.000000	0.000000
1	10000001	0.0	0.0	1.0	1.0	0.0	...	0.0	10223.421432	3945.661088
2	10000002	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	3571.095195

- 上傳結果

此次模型僅考量與消費類別最相關的總金額與總次數，分數呈現來到目前實驗個方法中的最高分。

Public	Private
0.6807	0.6797

4.5 計算 16 種消費類別的消費總金額與總次數 + 加入其他變數

- 模型

LGBMClassifier + `predict_proba`

- 方法描述

在上一個方法中，我們得到了明顯的進步，因此保留原本的資料格式，這次嘗試把其他資料加總加入更多變數，以及希望彌補 31000 個顧客的缺失資料中來增加準確率。有資料缺失的顧客我們在訓練集中，從時間最靠近的月份開始檢索，將其當作測試集資料，最後剩下約 4000 的 ID 是在第二年當中完全沒有消費記錄的人，除了顧客基本資料，顧客的信用卡紀錄資料全部以 0 實施填補。至此，全部 50 萬的 ID 都有可預測的資料。

- 上傳結果

此次方法給與模型更多變數以預測 Top3 消費類別，分數以是所有方法中最高的得分，比起使用統計 Top3 消費類別的方式來填補無資料的人的答案，以 0 填補加上基本資料，我們認為模型多少可以找出資料的相關性並給出較好的答案。

Public	Private
0.6843	0.6828

5. Results

第一次使用這樣大量的多分類題目資料，因為訓練、測試資料必須由自己建構，所以大部分的時間都耗在整理資料格式上，反而在模型預測的部分花費時間意外的少。最後提交答案為上方的方法 4.5，在競賽中獲得第 97 名，Private 與 Public 差距 0.0015。

TBrain AI實戰吧

Home

Competitions

Discussion

Datasets

Success Story


Overview

Leaderboard

Download Dataset

Submission History

93	塗錫錫大大大均	3	12	0.683298	1/6/2022 3:37:06 PM
94	T-Partner	4	1	0.683144	12/10/2021 10:35:57 AM
95	DART	6	30	0.683124	1/4/2022 7:20:27 PM
96	阿里山銀行	4	3	0.683099	12/23/2021 5:50:02 PM
97	巴斯克乳酪蛋糕	3	19	0.682762	1/6/2022 1:34:55 PM
98	東吳資科	5	1	0.682647	1/5/2022 9:08:25 PM
98	抱團求救組合	6	21	0.682647	1/6/2022 8:28:02 AM



859
參賽隊伍

開始 10/27/2021