巨量資料分析課程

玉山大數據競賽 報告書

H24074019黃科銓 H24076087高涵毅 H24071126葉嘉浤

國立成功大學 統計學系

一、問題描述

在本次玉山大數據競賽中，玉山銀行提供過去顧客信用卡消費歷史紀錄及基本資料，期待利用機器學習的方式，可以預測顧客最感興趣的三大消費類別排序，進而藉此做到精準行銷。本次預測目標為每位顧客在下個月(dt=25)消費總額中，針對重點16類消費類型商品，進行前三名的商品類別排序預測，並應避免預測出的類別重複名次，例如不接受前三名之中出現兩個名次是同一種類別的情形，最終預測結果則將以「NDCG@3」作為競賽評估指標。

二、資料處理

〔原始資料〕

本次資料集有53個欄位，約有3千3百多萬筆每月消費紀錄，共有50萬名不重複顧客，變數包含：

- 消費紀錄(43個變數)

  dt(消費月份，數值型)、chid(顧客編號，類別型)、shop_tag(消費類別，類別型)、txn_cnt(消費次數，數值型)、txn_amt(消費金額，數值型)、domestic_offline _cnt(國內實體通路消費次數，數值型)、domestic_online_cnt(國內線上通路消費次數，數值型)、overseas_offline_cnt(海外實體通路消費次數，數值型)、overseas _online_cnt(海外線上通路消費次數，數值型)、domestic_offline _amt_pct(國內實體通路消費金額佔比，數值型)、domestic_online_amt_pct(國內線上通路消費金額佔比，數值型)、overseas_offline_amt_pct(海外實體通路消費金額佔比，數值型)、overseas_online_amt_pct(海外線上通路消費金額佔比，數值型)、card_txn_cnt_1 ~15(卡片1到卡片14、與其他卡片的消費次數，皆為數值型)、card_txn_amt_pct_1~ 15(卡片1到卡片14、與其他卡片的消費金額占比，皆為數值型)

- 個人資料(10個變數)

  masts(婚姻狀態，類別型)、educd(學歷代碼，類別型)、trdtp(行業別，類別型)、naty(國籍，類別型)、poscd(職位別，類別型)、cuorg(客戶來源，類別型)、slam(正卡信用額度，數值型)、gender_code(性別代碼，類別型)、age(年紀，類別型)、primaty_card(正副卡標記，類別型)

〔資料清洗〕

- **異常值排除**

  在原始資料中，「消費金額」與「正卡信用額度」出現異常值，因此我們選擇將過大的數值改為資料的第99百分位數，以減少異常值的影響。

- **遺失值填補**

在原始資料中，屬於個人資訊的部分資料出現遺失值，尤其以「性別」與「年齡」這兩個變數有數萬個遺失值。在填補時使用基礎的方法，若是類別型資料，以眾數來填補，而數值型資料則使用平均值填補。

- **個人資訊變動**

進行資料清洗過程中，我們發現部分顧客的資訊會隨時間改變，例如行業別以不同數字做分類，處理資料後卻發現欄位出現小數點的狀況，後續觀察其餘變數，僅發現在行業別出現此狀況。

- **消費紀錄不完全**

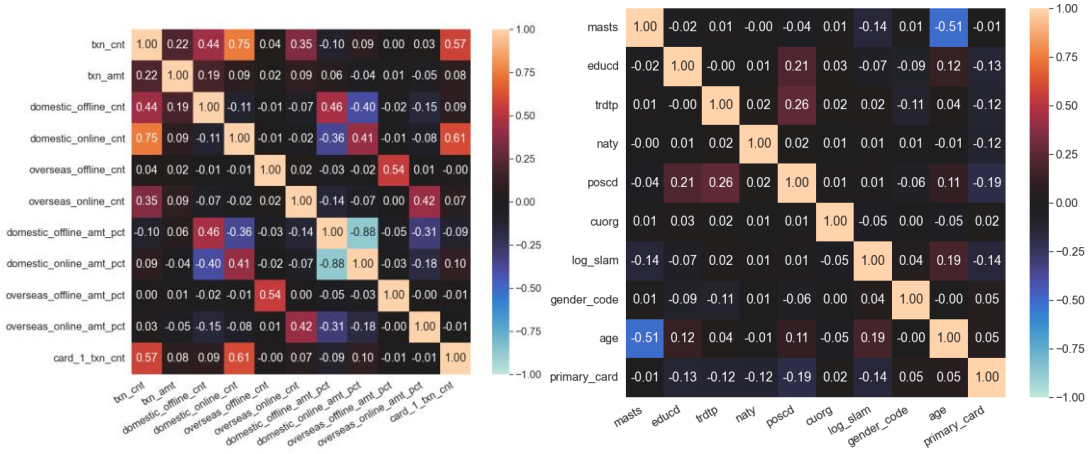
隨機取樣部分顧客後，發現多數顧客並非每月都會進行消費，甚至有顧客第一年完全沒有使用紀錄；此外消費紀錄的統計上也出現異常值，例如少部分顧客全部消費紀錄不到三筆，另外也有使用者的消費類別完全不屬於預測標籤的16類。針對上述問題我們分別以最新一次紀錄與統計眾人結果取代原始資料。

〔資料整理〕

我們將原始資料的消費比例轉換回消費金額，並以每一個顧客為單位，累計一年內的總消費紀錄，預測資料集可分成三個部分：今年消費資訊、個人資料、今年前三名消費類別，共有68個變數，預測目標則是明年的前三名消費類別，由於整理資料過程中，發現顧客之間的消費金額太懸殊（某些顧客消費金額過高，有些則為零），將造成後續模型建立的困難，因此採「取log」方式來縮小資料間差異，在資料切割方面上，以第一年的資料作為訓練集，第二年的資料作為驗證集，來檢驗模型好壞。

三、資料洞察

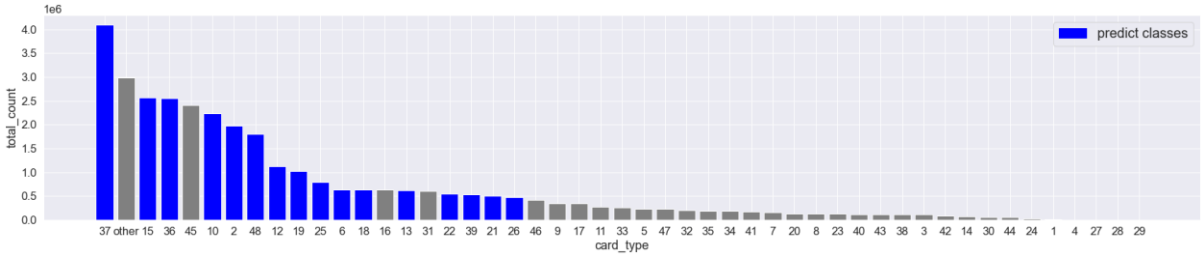
前半段使用簡單的圖表呈現資料探索的結果，後半段則以較直觀的統計方式做更細部的觀察。



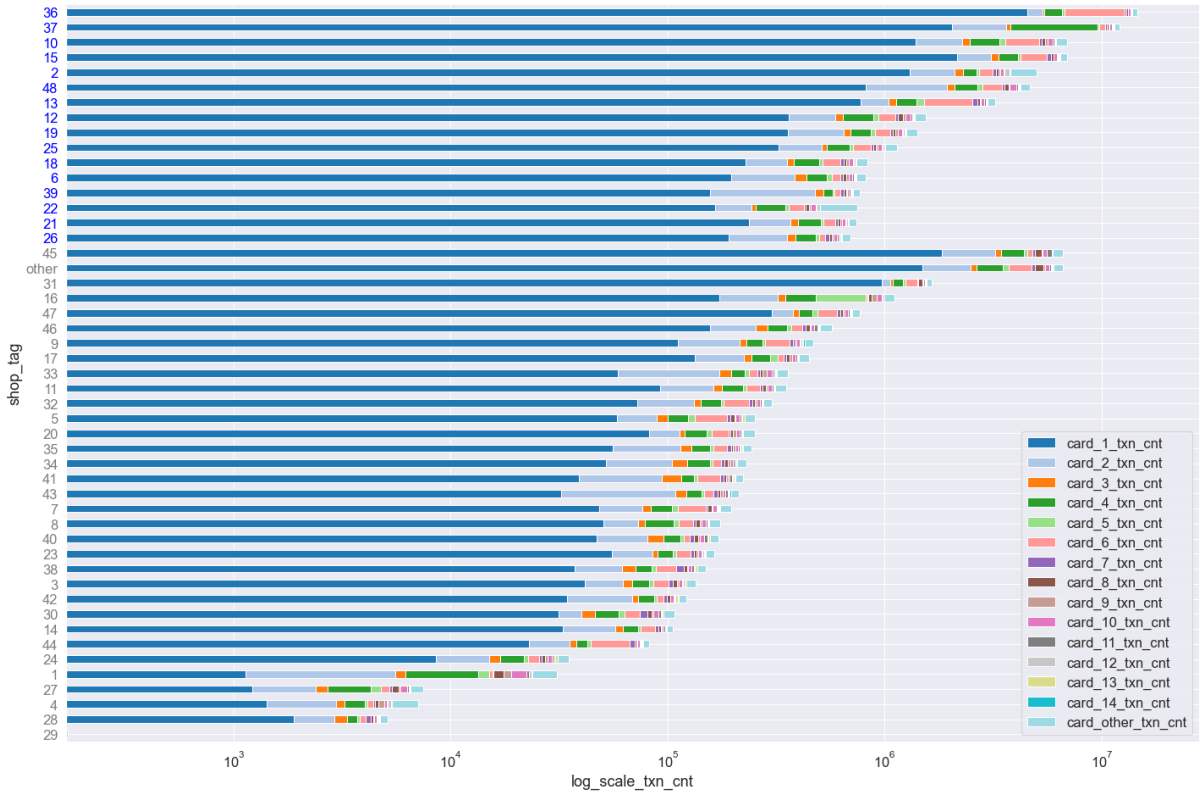
首先繪製數值型變數的熱力圖，上圖分別使用第24個月消費紀錄與個人資料的特徵繪製而成，可以發現消費總金額與國內線上消費總金額有0.75的高相關性，且在國內的實體消費比例與線上消費比例存在-0.88極高的負相關，可以解

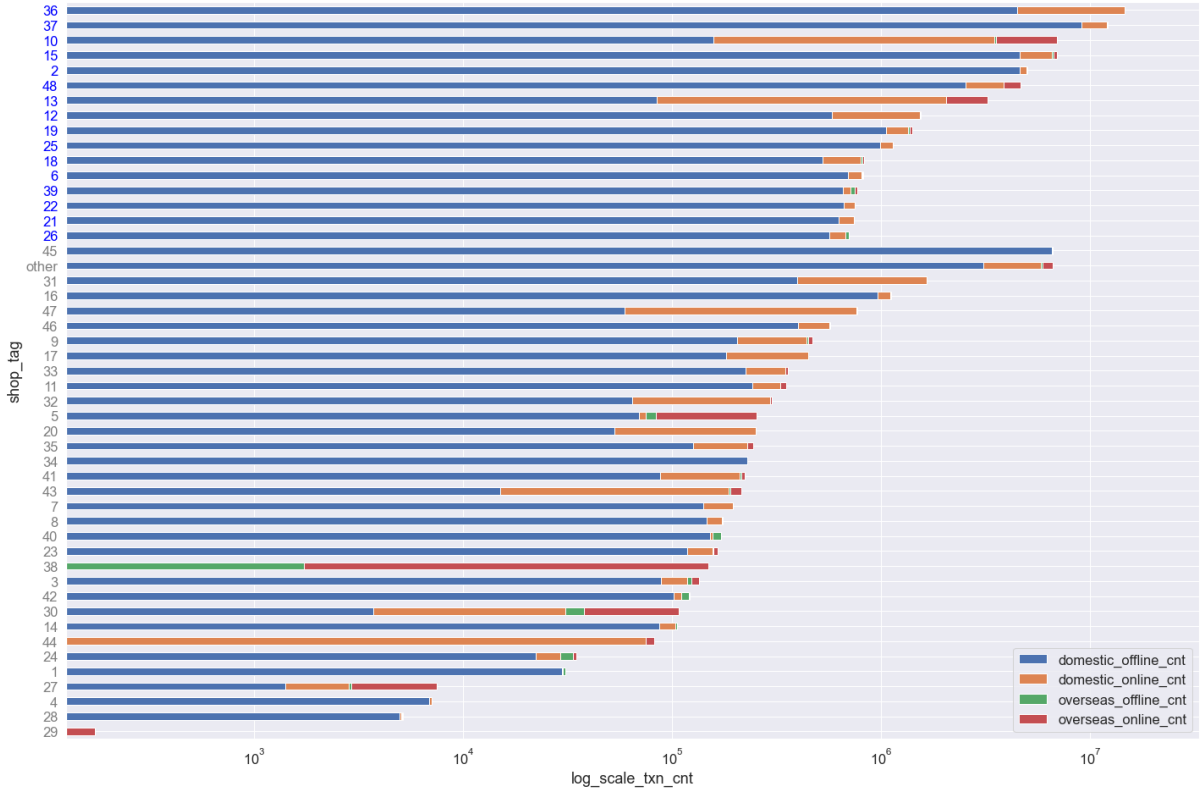
釋成持卡人主要在國內消費，且不會線上、線下同時消費的狀況；此外卡1同時與消費總金額、國內國內線上消費總金額有中高程度的相關，我們認為這三個變數存在有多重共線性問題。

接著使用長條圖觀察類別型變數的分布，下圖為累計所有消費類別發生次數的結果，塗為藍色的代表該標籤屬於16類其中之一。



最後繪製兩張堆疊長條圖，分別依據不同通路與不同卡片，對於消費類別進行次數統計，可以發現16類需預測的消費類別大多數仍以國內實體通路為主，且幾乎不存在國外實體的消費情況，此外所有消費類別主要皆以卡1進行購物，而其餘卡片則因不同類別暫居次要角色。





接下來我們依據系上課程所學，提出下列兩點的統計看法

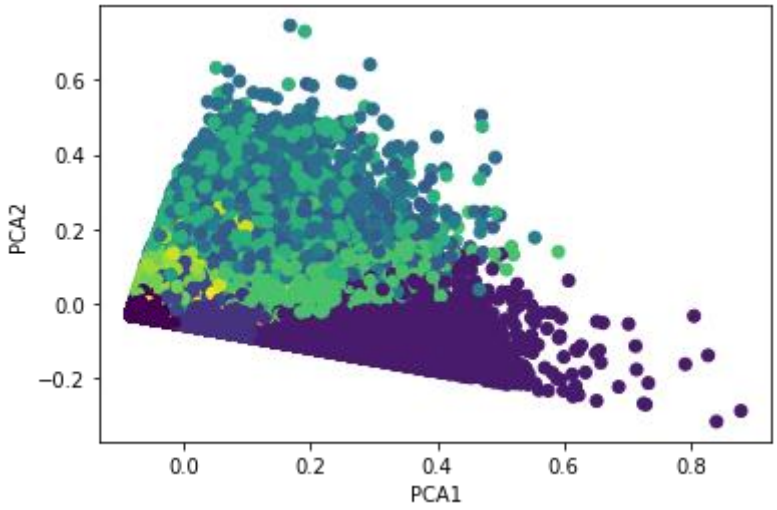
- 直覺法：不同方式的前三名消費類別

利用不同時間尺度、不同累計方式，我們可以統計出7種方法，首先以「消費總金額」來計算不同類別的總額，分別以第一年、第二年個別統計，可發現前三名消費類型皆為39、37、10、2，接著改以「消費總次數」來累計，分別以第一年、第二年個別統計，可發現前三名消費類型皆36、37、10、15，最後僅以「發生次數」來看，前三名則為37、15、36。在預測模型套用上，以「發生次數」效果最好，可見其他方法會受極端值影響，導致平均後的排名無法預測個別的前三名。

	消費總金額	第一年消費金額	第二年消費金額	消費總次數	第一年消費次數	第二年消費次數	發生次數
Top1	39	39	37	36	36	36	37
Top2	10	10	10	37	37	37	15
Top3	2	2	2	10	10	15	36

- 分群方法：以K-means方法分群

利用K-means分群的方式，先將原始資料的消費金額資訊正規化，去除極端值影響，再累計每個顧客在16類消費類別中的消費總金額，最終丟入分群模型中，第一次分5群，第二次分15群（參照下圖PCA1、PCA2的二維分群圖），希望透過此方法，可以減少繁重的資料量，資料相似度更為集中，讓預測模型得以更快速預測。



四、分析方法

針對這次消費類別前三名的預測目標，由於問題並非簡單的分類問題，還需要消費類別名次排序，因此我們提出五種可行想法來解決此問題，以下分別詳述：

方法一

僅以「消費類別top1」為預測目標，建立預測機器學習模型，例如隨機森林、Catboost、Xgboost、LightGBM等，並取出預測機率最高的前三名，作為top1、top2、top3，除此之外，也有嘗試利用K-means分群結果，訓練15個消費類別top1預測模型，來增加預測準確度。

方法二

以「消費類別top1、top2 、top3」為預測目標，分別建立個別top1、top2 、top3三個預測模型，並切同一個隨機seed的驗證集、訓練集，最終以「準確度」來作為模型好壞評估。

方法三

改成以「tag_top1、tag_top2 、tag_top3綁在一起」視為新類別標籤，例如top1為12，top2為15，top3為37，則新排序組合為〔12,15,37〕，作為預測特定序列前三名的類別預測模型，最終以「準確度」來作為驗證集預測評估指標。

方法四

運用「兼顧類別及排序預測」的LGBRanker來預測，設定一位使用者的所有消費紀錄作為一個group，以group為單位建立模型。可以設定用於訓練的group以及用於驗證的group數量，並以官方使用的NDCG作為驗證集的評估。

方法五

使用考慮時間效應的GRU模型，將此題目類比成NLP情感分析，直接對每一位使用者進行預測，計畫以前23個月的資料作為訓練集學習第24個月消費類別，透過此對應關係再以第2至24個月的資料推論第25個月的結果。

五、預測結果

以下預測結果呈現，將從五種方法來個別說明，並以最終競賽的NDCG@3來評估：

- **直覺填值**

預測說明	最終分數
全部顧客的**出現頻率**前三名	0.4327
個別顧客**消費總額**的前三名，若沒有前三名，則填大眾**消費總額**的前三名	0.5852
個別顧客**消費總額**的前三名，若沒有前三名，則填大眾**消費次數**的前三名	0.6258
個別顧客**消費總額**的前三名，若沒有前三名，則填大眾**消費出現頻率**的前三名	0.6416
第二年的個別顧客**消費總額**前三名，若沒有前三名，**隨機**從〔15,36,37,10,2〕類別取3個	0.6804

從上表的預測結果來看，若不用機器學習的方法，只純粹填前三名的消費類別，可以發現「消費出現頻率」>「消費次數」>「消費總額」，可見由於總額受到極端值影響，導致並非是好的個人模型預測，而比較全年資料與第二年資料，發現第二年資料預測的效果是最佳的，可見時間越近，更能最下一個月的預測。

- **方法一**

從下表的預測結果來看，可以發現LightGBM > Random Forest > Xgboost > Catboost，其中LightGBM不只是效果較好，模型訓練時間、可接受的資料集都較其他方法佳，我們也曾經想嘗試上課所學的Deep Forest模型，但最終卻因為資料極過大，而無法順利建立模型，另外，我們也有嘗試k-means分群方法來分別訓練模型，然而仍無法超過純粹填前三名的效果。

預測說明	最終分數
LightGBM	0.6722
Random Forest	0.671
Catboost	0.6493
Xgboost	0.6575
LightGBM(先K-means)	0.6634

- **方法二**

在此模型的預測上，最終效果不甚佳，除了top1預測模型在驗證集的結果，準確度可以達到0.7，其他top2預測模型僅達到0.4，top3預測模型更低於0.25，且三個模型預測出來的消費類別有很高機率會重複，會造成仍需要思考如何處理重複名次的問題。

- **方法三**

在此模型的預測上，遇到一個很大的障礙，就是合併後的類別種類過多（將近1000種），導致模型預測會花比較多時間、類別比重也會很懸殊，因此，有嘗試利用k-means分群結果，減少模型資料量，但仍是消費排序組合過多，處理時間仍偏長，仍需要思考如何加速處理時間，未來可以從「分群」來精進，好的分群結果，應可以讓組合變少一些。

- **方法四**

(一)將消費者依照ID分為五個super user，一個使用者作為一個group，在group內進行排序，最後取rank值最大的前三名作為答案

模型變數描述	最終分數
使用所有變數	0.433
將總金額乘上所有比例(通路、卡片的比例)、各通路次數、各卡片次數	0.483
使用特徵同上，但僅取dt=13~24的資料	0.5064
dt=13~24的資料、總金額取對數、各通路與各卡片次數、各通路與各卡片比例	0.506

(二)將消費者分為15群後，對這15個super user進行預測，將super user的所有消費紀錄排序後，取rank值最大的前三名

模型變數描述	最終分數
dt=13~24的資料、消費總次數、各個通路次數	0.4796
dt=13~24、總金額取對數、各通路與各卡片次數、各通路與各卡片比例	0.48

最終的分數，除了第一個使用所有變數的方法作為測試外，其他的分數約落在0.45~0.5左右。在第一個表格中，只使用第二年的結果會比使用兩年的結果稍微再好一點，而後面的特徵似乎還沒有對分數有很大的影響。將使用者分群形成super user後，分數似乎也和前面直接對使用者一一進行排序的結果接近。

直接對使用者一一進行排序，雖然能直接預測出50萬人的消費類別，但因為有些消費者的消費類別不在最後的16類、或是消費者的購買紀錄少於3筆，後續還需要我們進行人工的篩選與補值，需要不少的執行時間；而使用super user的概念進行預測，雖然可以避免上述需要人工寫答案的問題，但會有很大一群人的答案都是相同的，因此在選擇super user的數量上，或許是一個很重要的因素。

- **方法五**

在資料前處理的過程中遭遇較多難題，主要原因是資料過於龐大，即使挑選部分消費紀錄欄位作為特徵，但考慮24個月的時間區段仍讓特徵數量大於3000，因此在建立資料的過程花費許多時間，此外對於GPU的不了解讓後續建立模型時碰到技術上的障礙，導致在這個方法上並沒有在競賽中做出成果。

最終Private Leaderboard的分數為0.6804，本次競賽共有859組參賽團隊，我們排名為108名，大約落在前13%左右，比起Public Leaderboard的排名有更往前一些(分數=0.6809, 排名=15%)。

本次心得

統計大四 黃科銓

這次比賽最大的收穫，或者說是最大的挫折應該就是「資料太大了！」，我們這組便嘗試了很多方法，除了助教上次所教的PySpark，也有從網路找到巨量資料處理的好工具—Dask，不僅是讓資料處理效率變高，也可以做一些分群工具，像是這次分群的k-means就是用Dask內建的套件，另外，在模型的預測上，不斷「trial and error」，盡可能將每一個方法實現，雖然最終結果沒有很理想，但過程卻收穫滿滿，總結一下，這次經驗真的很可貴，我相信將可以幫助我未來再度遇到「巨量」時，可以不用再那麼不知所措了！

統計大四 葉嘉浤

Learning to Rank似乎是一個很新的議題，目前能在網路上找到的資訊還不多，雖然有模型可以使用，但過程的原理與結果，目前還無法很確定是如何進行的，對於我們在訓練模型與後續的預測時，花費不少時間研究。例如，模型的訓練方式有別於以往的模型，需要整理出可以使用的資料、模型內部的NDCG計算方式、預測出的分數不清楚要如何解讀。綜合上述的種種困難，我們在競賽的後期才成功使用這個模型，沒有對變數做轉換或是產生新的變數就直接預測，這可能是分數無法再提升的原因。

統計大四 高涵毅

再一次參加商業類型的數據分析競賽，但最終還是沒有妥善利用遞歸神經網路進行分析，主要原因除了資料的筆數過於龐大導致處理上的困難，另外針對Rank Data的預測因為之前沒有接觸，在關鍵字的找尋上一直無法與RNN這類模型找到很好的連結，不過在前期資料探索的鑽研，這部分整合了過去三年系上所學，算是一套非常完整的分析流程，確實幫助我們更快掌握巨量資料的分布狀況，希望未來有機會可以將統計方法與AI模型完美的搭配起來。