國立臺灣大學管理學院創業創新管理碩士在職專班

碩士論文

Entrepreneurship and Innovation MBA Program in Business Administration College of Management

National Taiwan University

Master Thesis

使用者行為數據分析：

以線上學習歷程紀錄分群為例

User Behavior Data Analysis: A Case Study of Clustering Online Learning History Records

卓東昇

Tung-Sheng Cho

指導教授：孔令傑 博士

Advisor: Ling-Chieh Kung, Ph.D.

中華民國111年8月

August, 2022

doi:10.6342/NTU202202818

致謝

首先，感謝孔令傑老師這段時間的教導，從資料分析的課程到論文題目的選定，以及論文撰寫期間的架構討論與細節修正，使我能順利完成碩士學位論文並完成口試，讓我的碩士生涯畫上完美的休止符。同時也感謝陳聿宏教授與余峻瑜教授願意參加我的口試並給予寶貴的意見，讓我獲益良多，也使論文更加完善。最後感謝瞿志豪老師，如果沒有瞿老師的介紹，或許就會因此和孔令傑老師錯過，錯過最好的指導教授。

另外也感謝與我一同完成本個案專案的許晉芳、劉楚軒與朱家慧三位同學，沒有各位的參與，斷然不可能在兩個月內完成專案，並提供本研究完整的素材。在此感謝各位對本論文的貢獻。

碩士的生涯終於也走進尾聲，回首這段時間，無論是工作、家庭還是學校都充滿了各式各樣的挑戰。感謝老婆與孩子們的體諒，接受三天兩頭不在家，假日還常常缺席重要活動的我，支持並鼓勵我完成學位，將此論文獻給你們。

doi:10.6342/NTU202202818

中文摘要

數據分析已是現代商業不可或缺的一環，企業也開始蒐集使用者行為數據以作為其商業決策的基礎。但由於企業難以建立兼備產業知識與數據分析能力的團隊，亦難以評估導入效益，導致企業缺乏導入數據分析的意願，即便勉強執行也常導入不順。

本文以線上教學平台導入分群模型作為產品開發基礎為例，詳列企業導入數據分析的過程與挑戰，以作為企業規劃的參考。本文詳細記錄如何使用平台學生端的線上行為數據，透過實際進行數據清理篩除不必要與錯漏的資料，並以敘述性統計方法進行分析後選擇並建立學生特徵值，最後使用簡單統計量分群方法成功建立分群模型，並於教師端與企業端皆取得正面的回饋，成為下一階段產品開發的基礎。

此個案研究驗證了使用者行為資料分析具備提高企業價值的能力，並透過此個案研究記錄了完整資料分析流程與相關實作細節，最後提出企業初期導入及擴大實行的建議。

關鍵字：數據分析、使用者行為、使用者分群、線上教學平台

doi:10.6342/NTU202202818

Abstract

Data analysis has become the most important function of a modern business, and companies also begin collecting user behavior data for their business decisions. However, it is too difficult to recruit a team with both domain knowledge and data analysis capabilities, so the companies are hard to evaluate the benefits of introducing data analysis. Because the companies have few interests to introduce data analysis, and even if they are trying to do so, they often fail.

This article takes an online education platform as an example of implementing data analysis in a company. The results of this case were eventually turned into the foundation of product requirements for their future iteration. The article records in detail how to use the online behavior data of students on the platform, removes unnecessary and erroneous data through the data cleaning process, and analyzes it with descriptive statistical methods to select the student features. At last, a clustering model was successfully built by using simple statistical values, and the model got positive feedback from both the teacher and the enterprise.

This case study assures that user behavior data analysis is able to improve business value, and records the complete process and implementation details for reference.

doi:10.6342/NTU202202818

Finally, the case study concludes with suggestions for the early implementation and the expansion when enterprises introduce data analysis to their business processes.

Keywords: data analysis, user behavior, user clustering, on-line education platform

doi:10.6342/NTU202202818

目錄

doi:10.6342/NTU202202818

doi:10.6342/NTU202202818

圖目錄

doi:10.6342/NTU202202818

doi:10.6342/NTU202202818

表目錄

doi:10.6342/NTU202202818

第一章 緒論

1.1 研究背景與動機

從網路時代開始，許多企業透過網路對使用者提供服務，也因此累積了大量的使用者行為數據，這些數據可以讓企業建立分群模型，在現今強調分眾行銷的背景下，應用極為廣泛，常見如：

1. 金融業分析客戶信用狀況給予貸款額度、理財用戶的金融商品推薦。
2. 零售業用於區分會員等級、商品推薦、精準給予商品廣告。
3. 線上串流產業區分會員以推薦影片。
4. 社群平台透過分群推薦使用者感興趣的內容或給予不同類型廣告。

這其中也包含了線上學習平台，如 Coursera 等 MOOCs 的興起，同樣地累積了許多數據，也有許多人研究如何使用平台的使用者行為數據來建立分群模型，來達成分眾化的教學。

導入數據分析的好處顯而易見，企業也爭相導入。根據科技新報(TechNews)於 2017 年發布之「2017 個人/企業科技大趨勢調查」，有 59%的受訪者認為，數據分析在企業應用具有潛力，並認為有助於決策，並且有 73.6%的受訪者預計於三年內導入。然而四年後，2021 年 Forrester Consulting 的報告卻指出，台灣仍有 59%的企業尚未導入數據分析，這表示企業在導入數據分析時必然遇上了一些問題。而數據分析導入的問題主要可以分為三個方向：

1. 企業數據不足：據 2021 年 Forrester Consulting 的報告，有 78%的企業表示，企業擁有的數據不足。
2. 企業技術能力不足：同樣在 2021 年 Forrester Consulting 的報告，導入數據分析的企業中，有高達 67%的企業表示，數據產生的速度遠超出企業所能分析的速度。

doi:10.6342/NTU202202818

3. 企業需求不明確： IDC 軟體暨服務市場分析師蔡宜秀(2015)曾表示，企業最常遇到的數據問題是確認需求。

企業技術能力與需求環環相扣，企業若要準確地提出需求，就需要實際了解企業內部運作方法並具備豐富產業經驗的人，然而這種人往往不具備技術能力，不知道技術的可行邊界；即便企業聘請了數據分析人員，數據分析人員通常也不具備相關的產業知識，無法了解企業的實際需求。在這樣的情況下，企業端甚至無法蒐集有效數據，一旦數據不足，即便企業同時具備數據分析能力與明確需求也無法進行分析。因此，我們可以歸納出，若要確保數據分析能順利導入，擁有充足的數據，並能協同具備產業知識的企業端窗口以及具備資料處理能力的數據分析師，將是導入成功的關鍵。

本研究透過實際與營運中的線上學習平台合作，由企業端準備數據，透過與具備產業知識的企業端窗口溝通與協調，協助企業端導入數據分析並建立分群模型，本研究試圖將此次導入數據分析的經驗，將此次導入的過程詳細記錄下來，希望能帶給企業端數據分析部門產業實務與數據分析互相合作的相關經驗，讓未來數據分析師在企業內部推廣數據分析能更加順利。

1.2 研究目的

在本研究中，我們將以一線上學習平台為個案研究對象，透過詳細地紀錄如何以數據分析流程解決該平台的使用者分群問題的需求，讓數據分析部門能依此個案經驗，建立符合業務端需求的數據分析流程，並預先了解過程中可能面對的問題。透過個案研究方式，本研究欲達成的目的如下：

1. 建立線上學習平台之學生分群數據分析架構與流程。
2. 透過個案展示如何針對企業需求進行數據分析。
3. 導入數據分析對於本個案之企業帶來哪些價值。

doi:10.6342/NTU202202818

1.3 研究方法與流程

本研究使用個案研究法。透過實證探討記錄我們如何使用個案企業資料，依資料清理、敘述性統計、選擇能區隔使用者的特徵值、建立使用者特徵值表、建立分群模型等流程，在分群模型建立後交付個案企業產品經理，並將分群結果給予教師參考，如圖 1 所示。

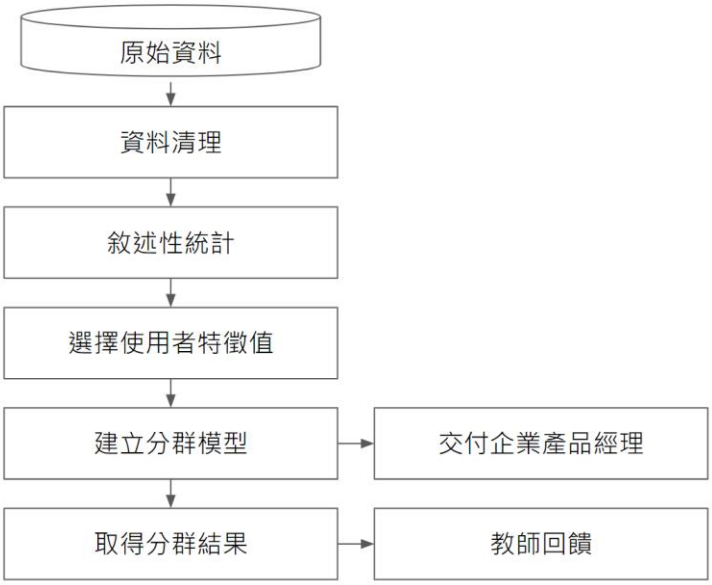


圖 1、研究流程

doi:10.6342/NTU202202818

第二章 文獻探討

本研究的主要目的在於如何針對企業需求進行數據分析的導入，並透過個案研究，紀錄我們如何從線上學習平台使用者行為數據製作分群模型的過程。因此我們關注二類文獻：使用者分群在各產業的應用及線上學習平台的使用者分群研究。

2.1 使用者分群於各產業之應用

數據分析於商業領域之應用已行之有年，對使用者進行分群是數據分析中的重要的應用領域。在網際網路剛剛興起的年代，就已經有透過演算法分析使用者行為數據並建立使用者輪廓來協助企業進行精準投放廣告、推薦、或讓網站呈現個人化的相關研究(Xie, 2001)。

使用者分群常運用於零售業，精準行銷是當前零售業數位行銷的重點，零售業可以透過自己所擁有的顧客資料進行數據分析，找出顧客潛在的喜好並為顧客提供更加精準的產品。其中會員制的運用可以幫助企業收集顧客的使用者行為數據，使用 RFM 模型對顧客進行分群，RFM 模型是由 George Cullinan(1961)提出，RFM 分別代表最近一次消費(Recency)、消費頻率(Frequency)、與消費金額(Monetary)。取得消費者的 RFM 後，可以直接以中位數或平均數等簡單統計量來進行分群，最簡單的 RFM 模型就可以以此方法分為八個分群；也可以在每個維度都切成五份，如此就會有 125 個分群。也有結合 RFM 與 K-means 分群的方法(Ching-Hsue Cheng, 2009)，在使用 RFM 方法抽出顧客的特徵值後，再使用 K-means 分群，最後使用決策樹或 RS Theory 進行分群解釋(Daqing Chen, 2012)。

由零售業的使用者分群研究來看，使用者分群的作法大抵可以歸納為幾個步

doi:10.6342/NTU202202818

驟：

1. 收集使用者資料。
2. 抽取使用者特徵值(在零售業常用 RFM 及購買過的商品)。
3. 選擇分群模型(使用簡單統計量分群方法、K-means、或其他方法)。
4. 解釋分群模型的運作邏輯(若使用 K-means 可以搭配決策樹)。

其中 K-means 的分群數量需要被預先決定，會於第 2.3 節再做討論。

金融業與社群平台也遵循相同的流程，以金融業為例，可以單純使用 RFM 取得客戶特徵值或使用複雜資料(Zakrzewska, 2005)，取得客戶資料後，再使用 K-means 或其他演算法如 DBSCAN 進行分群；而社群平台則可以將社交圖譜轉化為使用者的特徵值，再進行分群(Ahmed Alsayat, 2016)。此一框架在線上學習平台的研究也依然適用，線上學習平台的相關研究整理於第 2.2 節。

2.2 使用者分群於線上學習平台之相關研究

Lin-siegler(2016)認為對學生施予不同教育方式可以改變學生的學習歷程並有可能提高學生的表現。對應線上學習平台，則可以透過蒐集學生在平台上的行為數據來對學生進行分群，並針對不同分群的學生改變平台的行為來提升學習效果(Mojarad, 2018)。絕大多數的線上學習平台的使用者分群都限縮在線上學習平台來影響學生，但也有透過學生的線上學習資料建立分群模型後，同時改變線上及線下教學行為並比較差異的研究(Šarić-Grgić, 2020)。使用的分群方法多數為 K-means，分群數設定在 2 到 4 之間，但也有部分研究使用 Mean shift 或最大期望演算法，在使用 K-means 的場合中，可以透過決策樹來取得對分群方法的解釋(Šarić-Grgić, 2020)。

doi:10.6342/NTU202202818

目前線上學習平台的相關研究著重於分群模型的好壞，而並未記錄如何與企業端互動以建立分群模型，所得到的分群結果也較沒有對外溝通使用，我們後續的個案研究將會著重在記錄與企業端如何互動，並研究如何對外溝通分群結果。

2.3 K-means 與 RFM 比較

綜合前兩節所提到的分群方法，考量其易用性，選擇 K-means 與 RFM 模型做為最後的比較。

K-means 分群演算法需要預先輸入分群數量，常見使用其他相似演算法取代，或使用 SOM 來預先取得合適的分群數量(Dogan, 2013)。離群值的問題也可以透過 preprocessing 來處理 (Patel, 2011)，最大的問題是 K-means 的分群邏輯不易解釋。在研究中，使用決策樹或 RS Theory 等可解釋性的機器學習方法做對映模型來取得分群邏輯，但是也僅只是接近而非完整的分群邏輯，因此實務上，K-means 常用在不需要與外部溝通分群邏輯的場合，如對應不同分群的使用者發送電子報等。

RFM 模型架構簡單，容易解釋並微調分群邏輯，但是特徵值與分群界線的選取會影響分群的效果，且特徵值如為類別資料，會因此較難決定分群的分界點，適合需要對外部溝通分群邏輯的場合，如於行銷活動制定行銷規則等。

K-means 與 RFM 模型比較如表 1：

表 1、K-means 與 RFM 模型比較

分群演算法	K-means	RFM 模型
演算法簡介	最常使用的分群演算法。	架構簡單，且容易使用
優點	簡單易操作，且沒有輸入資料的限制。	使用簡單統計量做分群依據，分群邏輯易於解釋，且分

doi:10.6342/NTU202202818

		群自由，易於對分群邏輯微調。
缺點	1. 需要預先輸入分群數量 2. 若有離群值會導致分群模型表現不好 3. 分群邏輯不易解釋	1. 特徵值與分群界線的選取會影響分群效果 2. 面對類別資料，難以決定分群依據

本個案將會比較此兩種分群模型，並擇一使用。

doi:10.6342/NTU202202818

第三章 實證探討

3.1 個案描述

此個案之企業為一線上教學平台，線上教學平台是透過網際網路遞送教學物件之軟體或網站，教學物件可能是影片、文字、或是錄音等。以下分為三節，分別介紹此教學平台、該平台提供之使用者行為資料以及該企業期望該行為資料庫能轉化的產品標的。

3.1.1 個案教學平台概述

本個案企業為一針對國小、國中與高中銷售之線上教學平台，企業希望透過此平台的各種功能提高學生閱讀素養，並且為企業之主要價值主張。此平台其中一項產品會定期發放學生閱讀測驗，後稱為「任務」。任務包含：一篇閱讀時間約為 5 分鐘長之文章與 5 道對應該文章的題目。任務每周會分發二次，每次包含兩個任務，每週共有四個任務。國小國中高中會分別收到不同的任務，一般任務只在學期時發放，寒假與暑假則會發放與學期中一般任務不同的特殊任務。

任務的文章依內容主題不同，分為十類：世界櫥窗、人文史地、人物、文化、文學、生活知識、社會科學、自然科學、說明文件、議題等，又稱為文本領域；任務的題目也依題型的不同分為三類：連續、連續含圖表、非連續等，又稱為題目類型。

該產品使用者分為二群，分別為學生與教師。產品主要使用者為學生，系統會紀錄學生的班級與座號，並在作答過程中產生相對應的使用紀錄資料。而教師除了可以檢視閱讀任務的文章與題目外，也能看見任務的其他資訊如難度與正確答案等。此外，教師可以透過檢視自己班級學生的任務作答狀況與作答記錄來修

doi:10.6342/NTU202202818

正使用此閱讀測驗工具的方式或改變自己的教學方法。

根據個案企業提供的資料顯示，此產品目前主要提供花蓮的國小國中與全台的私立國中小與高中使用。

3.1.2 使用者資料敘述

學生收到任務後，可以在平台上進行作答，作答時間為一小時，若超過一小時則自動送出答案，並視為學生第一次答題。答題後若五題皆正確則為完成任務，任務未完成前可以重複答題。學生在平台上答題後會留下紀錄，包含：初次答題數、初次答題正確數、首次答題使用時間、首次完成任務時間點等，於第 3.2.2 節有詳細定義。

3.1.3 企業期望之資料應用方法與產品標的

透過分析與使用學生在平台產生之行為數據，可以依據某些指標來對若干學生進行分群。例如：我們可以同時分析全校的學生，依據全校全部任務的首次答題時間的中位數將全校分成兩群；或是我們抽取某一班的學生，依是否完成所有任務分成兩群。

個案企業預期，若分群結果具有邏輯上的意義，透過將此分群結果提供給教師，教師就能運用分群結果對不同群的學生採取不同的措施，舉例來說，教師可以對任務完成度較低的學生嚴格追蹤並要求在時限內完成，或是針對某些領域表現較差的學生進行特別輔導。此功能可以使教師與平台的互動增加，進而提高該教學平台對教師的價值。若能因此提高學生閱讀素養，也有助於平台的銷售並提升其品牌價值。

在訪談企業產品經理並達成共識後，我們的分群模型將鎖定此閱讀測驗產品的一般任務上，並依學生使用行為數據建立分群模型，最後再實際產生分群結果

doi:10.6342/NTU202202818

供企業與教師端驗證，以確認分群方法是否具備說服力與可行性。

3.2 資料清理

從個案企業取得二筆資料庫共 4,179,163 筆，分別來自花蓮國中小 1,636,852 筆與私立學校 2,542,311 筆，時間從民國 108 年到民國 111 年，若以學期紀錄，則為 107-2 到 110-2，後續時間計算以學期為主。

3.2.1 資料欄位定義

平台端每次新增任務，資料庫會對應每一個收到任務的使用者產生新的一筆新資料，每筆資料會記錄：

1. 該任務相關資訊。
2. 該使用者相關資訊，。
3. 該使用者對應此任務的作答狀態。

資料欄位之組成及其新增與更新時間點如圖 2。

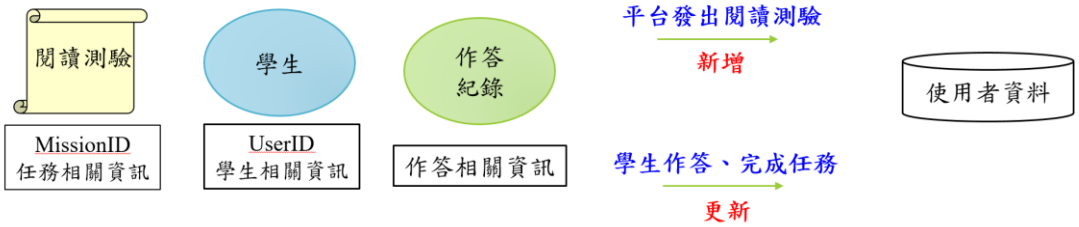


圖 2、資料欄位之組成及其新增與更新時間點

舉例來說，平台端增加一項任務，該任務共發給 1000 個學生，則會產生 1000 筆資料紀錄每個學生對應該任務的作答狀態。資料欄位中的每筆資料都包含 24 個欄位，資料欄位名稱與定義如表 2 至表 4。

doi:10.6342/NTU202202818

在學生首次作答後，資料欄位中關於首次答題的資料欄位將被更新，如表 4 編號 17、18、19、20、21、22 的資料欄位；在學生完成任務前，資料僅會更新目前任務狀態，也就是表 2 的編號 19；直到學生完成任務後，紀錄任務完成，更新表 4 編號 19 與 23。

表 2、資料欄位與定義(任務相關)

編號	欄位名稱	定義
1	Product	產品別，包含目標產品與其他產品名稱。
2	YearOfMisson	任務派發年度。
3	MonthOfMisson	任務派發月份。
4	SemesterYear	任務派發學期。
5	MissionName	任務名稱。
6	MissionType	任務類別，分為一般任務與特別任務。
7	Difficulty	文本難易度，分為三個等級，難、中、易。
8	Field	文本領域，分為十個領域。
9	Type	題目類型，分為三種類型。
10	MissionTime	任務派發時間。
11	MissionID	任務 ID。

表 3、資料欄位與定義(使用者相關)

12	Grade	使用者為學生，學生年級；教師為空值。
13	userID	使用者 ID。
14	Position	使用者身分，分為教師或學生。
15	ClassID	使用者所在班級。

doi:10.6342/NTU202202818

16	SeatNumber	使用者為學生，學生座號；教師為空值。

表 4、資料欄位與定義(作答狀況相關)

17	CorrectCount	使用者首次作答的答對題數
18	AnswerCount	使用者首次作答的答題題數。
19	MissonStatus	目前任務狀態，分別為： 1. 收到任務：使用者收到任務。 2. 開始任務：使用者開始做答。 3. 完成任務：使用者完成任務。 4. 暫停任務：使用者開始做答，但未完成任務。 5. 錯誤任務：其他例外狀況。
20	CorrectRate	學生第一次作答的正確率，為前述 CorrectCount / 5。
21	FinishTime	使用者首次完成任務時間。
22	AnswerDuration	使用者首次答題使用時間。
23	isCompleted	任務是否已完成，為 YES / NO
24	UpdateTime	此資料最後更新時間。

3.2.2 預計資料清理流程

取得原始資料後，需對資料進行清理並逐步建立操作資料，考量最後成品應為學生分群模型，因此原始資料至少應去除非學生的使用者資料、錯誤的資料以及有部分使用者缺乏資料的欄位，預計過程如圖 3 所示：

doi:10.6342/NTU202202818

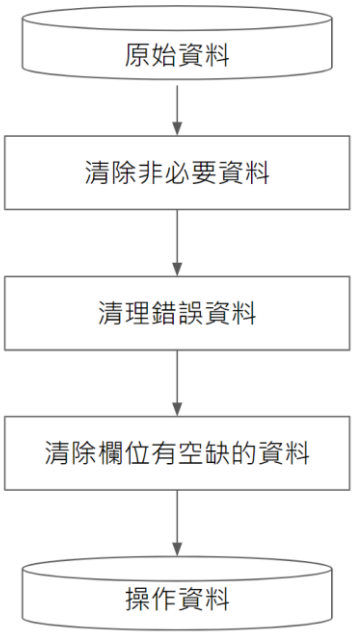


圖 3、預計資料清理流程

3.2.3 敘述性統計

在清理過程中需要時時對資料進行敘述性分析，可以藉此與企業產品經理確認資料是否在處理過程中發生錯誤，並維持和企業端認知一致；此外，透過觀察資料型態，能夠從各方面確認資料的分布型態，從而得到後續資料處理的洞見。圖 4 至圖 11 為花蓮與私立學校在分析過程中的敘述性統計示意圖。其中，y 軸表示「資料筆數(任務總數)」；領域表示閱讀測驗文本的內容主題，分為十種；類型表示閱讀測驗的題目類型，分為三種。

doi:10.6342/NTU202202818

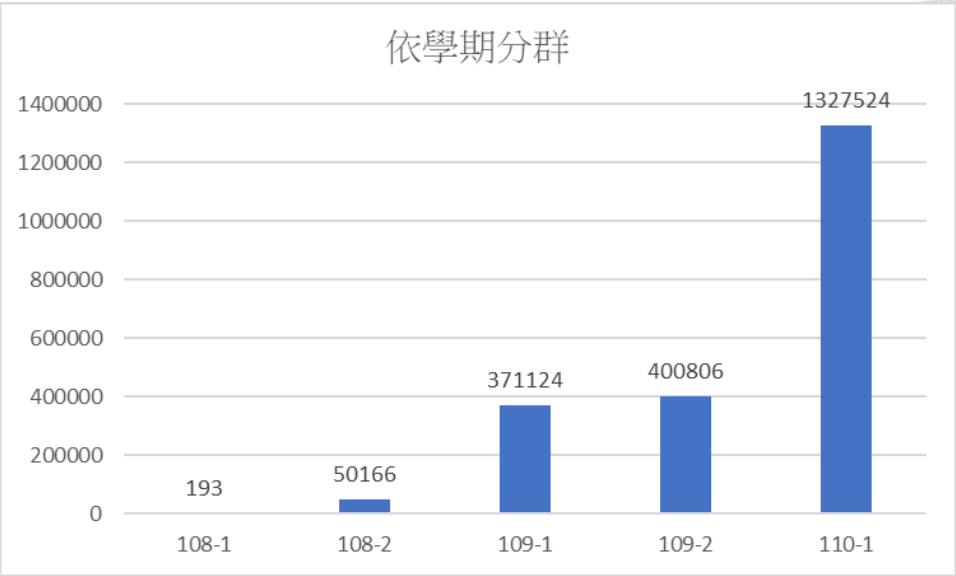


圖 4、私立學校資料分布示意圖(依學期)

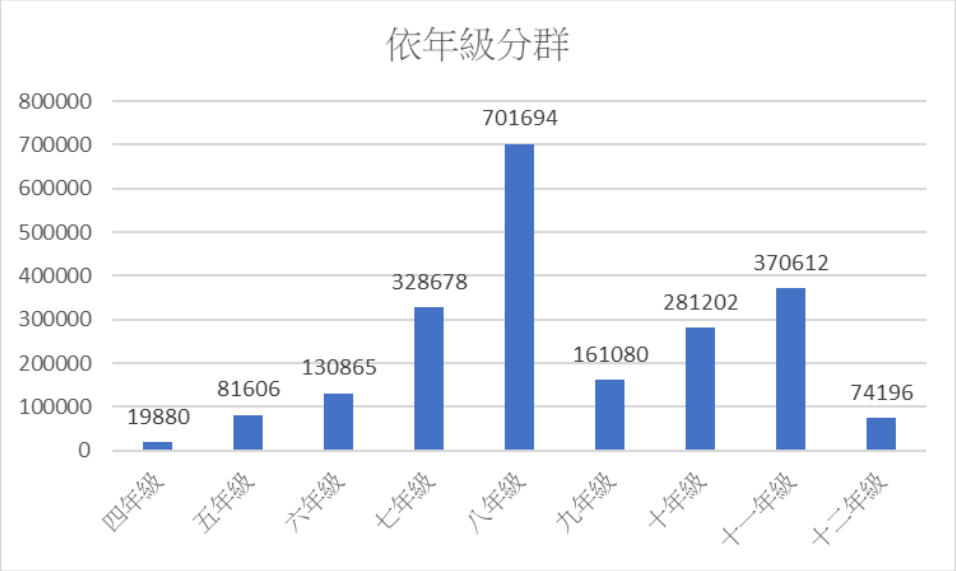


圖 5、私立學校資料分布示意圖(依年級)

doi:10.6342/NTU202202818

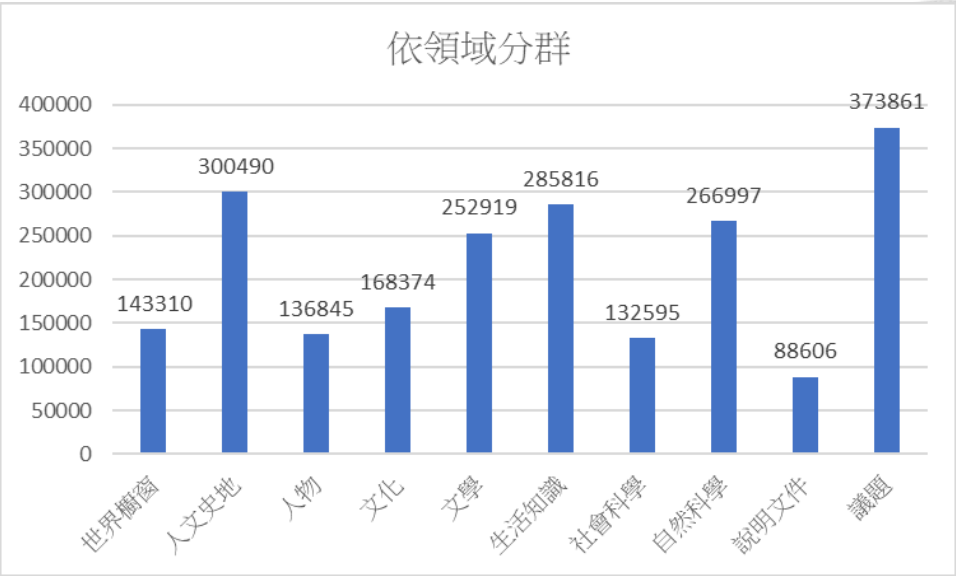


圖 6、私立學校資料分布示意圖(依文本領域)

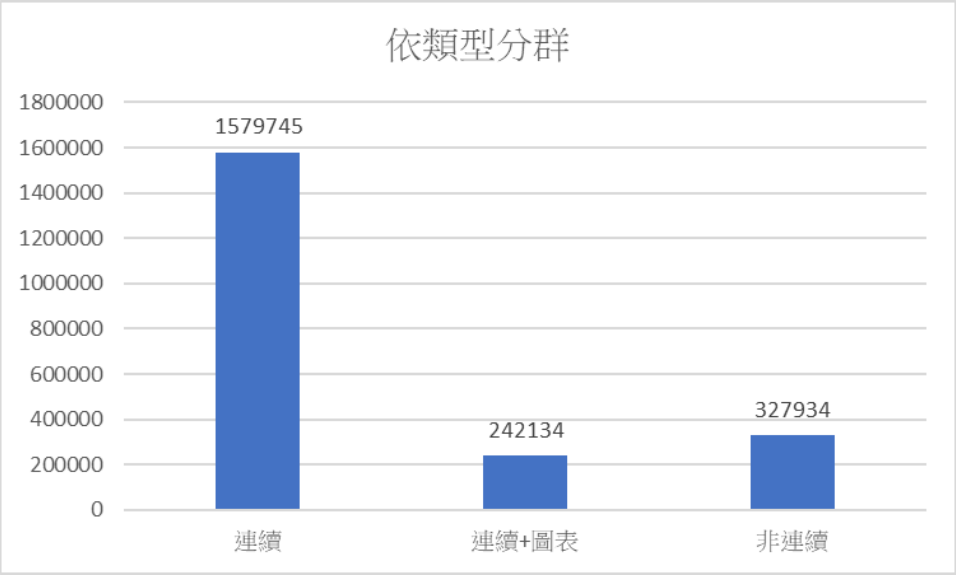


圖 7、私立學校資料分布示意圖(依題目類型)

doi:10.6342/NTU202202818

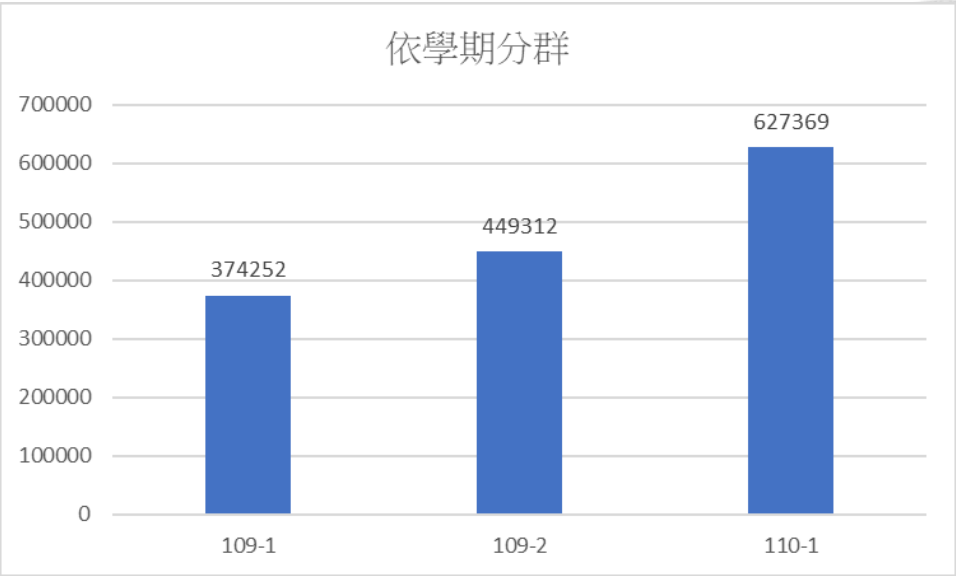


圖 8、花蓮學校資料分布示意圖(依學期)

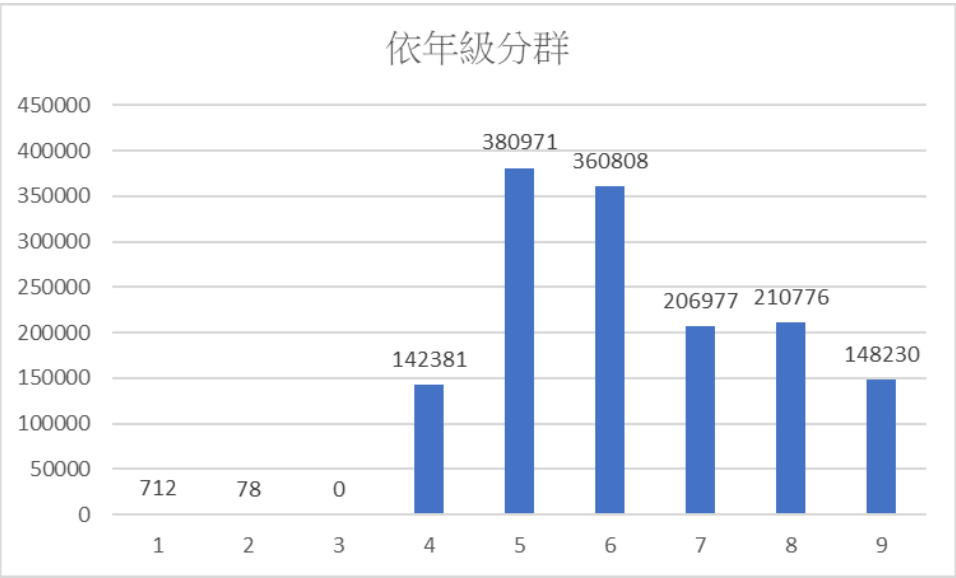


圖 9、花蓮學校資料分布示意圖(依年級)

doi:10.6342/NTU202202818

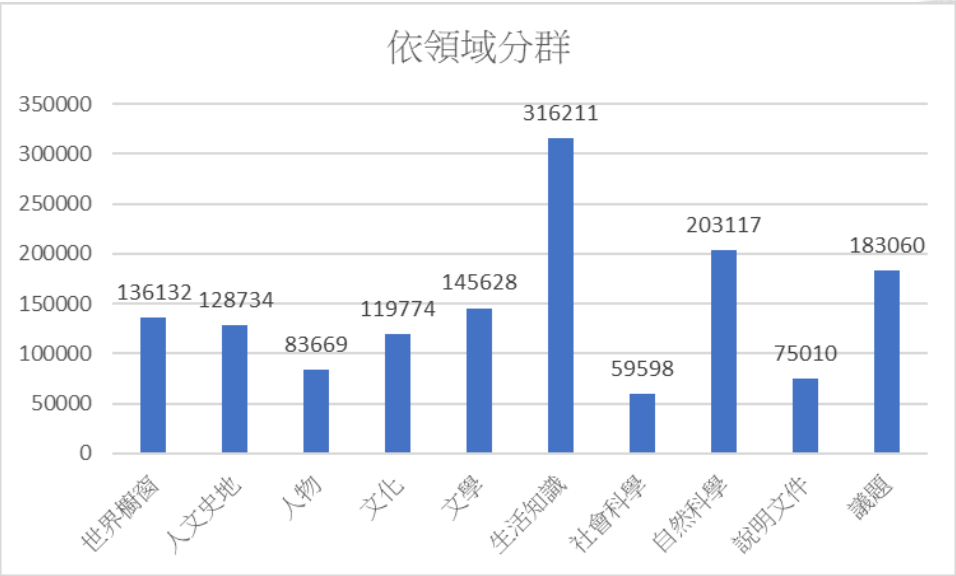


圖 10、花蓮學校資料分布示意圖(依文本領域)

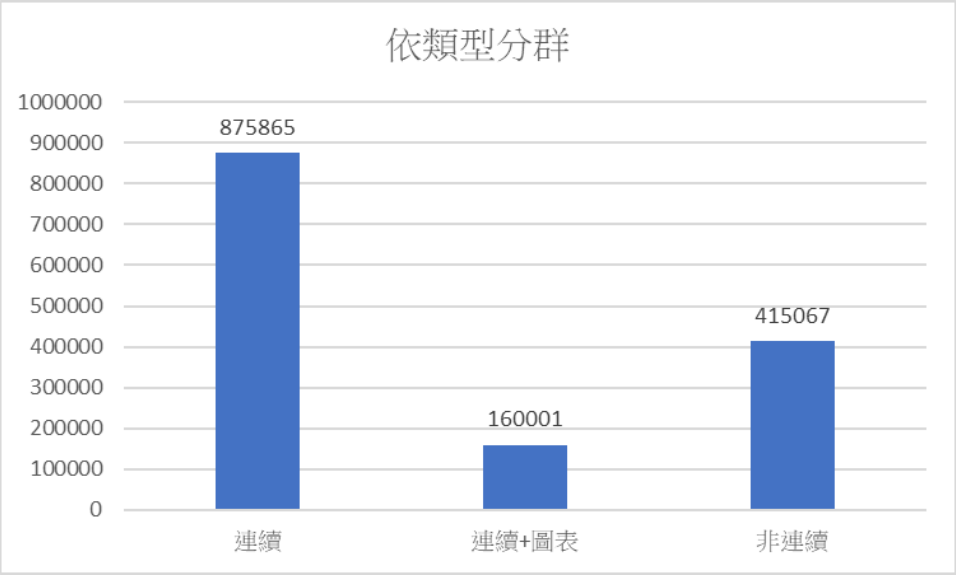


圖 11、花蓮學校資料分布示意圖(依題目類型)

3.2.4 實際清理流程

實際過程紀錄如下：

1 清除非必要資料：如企業端產品經理要求，我們僅保留學生在閱讀測驗產品的一般任務使用資料，因此我們依序刪除以下資料：

1.1 學生以外的使用者資料，僅保留 Position 欄位為學生的資料，在這裡我們清除了老師與企業測試人員的資料。

doi:10.6342/NTU202202818

1.2 非閱讀測驗產品的使用資料，僅保留 Product 欄位為閱讀測驗產品的資料，在這裡我們清除了個案企業另兩項產品的資料。

1.3 任務類型非一般任務的資料，僅保留 MissionType 欄位為一般任務的資料，這裡清除的主要是特別任務的資料。

2 清除錯誤資料：我們列出幾項邏輯或有衝突之處，在分析資料後列表向企業產品經理提出，由企業產品經理確認資料處理方法：

2.1 單一學期中，ClassID 和 SeatNumber 應對應到單一的 userID，但實際列出後，私立學校的資料中有 9 個 userID 有超過一組 ClassID 和 SeatNumber；花蓮學校的資料中則有 23 個 userID 超過一組 ClassID 和 SeatNumber，經確認後刪除 64 筆資料。

2.2 MissionStatus = 3 應與 isComplete 的結果一致，但經檢測有 4062 筆資料不符合此條件，經確認後刪除 isComplete 欄位，以 MissionStatus = 3 做為任務完成的唯一條件。

2.3 AnswerCount 不應該大於 5，因為總題數只有五題，使用者不可能回答超過 5 題，共刪除 3 筆資料。

2.4 有部分欄位內資料不在定義的資料範圍內，例如 Field 欄位內的值非定義的九種領域別、Grade 出現非年級的字串，經確認後全數刪除。

3 清除有空缺資料的欄位：有極少數欄位為空值，與企業產品經理討論後，認為不影響分群模型，因此選擇刪除整筆資料。

4 清除特定資料：並未在初始預計的資料清理流程中，經與產品經理討論後，刪除：

4.1 國小一、二、三年級資料，原因為資料過少無法單獨建立分群模型，若要建立分群模型必須併入國小四、五、六年級，經討論後刪除。

4.2 特定學期資料：企業產品經理表示在特定學期時，該產品正在進行測試，因此有混入大量非學生產生的資料，但沒有分辨的方法，經討論後，直

doi:10.6342/NTU202202818

接刪除該特定學期資料。

4.3 剩下的資料定義為操作資料，詳細結果如第 3.2.5 節。

4.4 實際流程如圖 12。

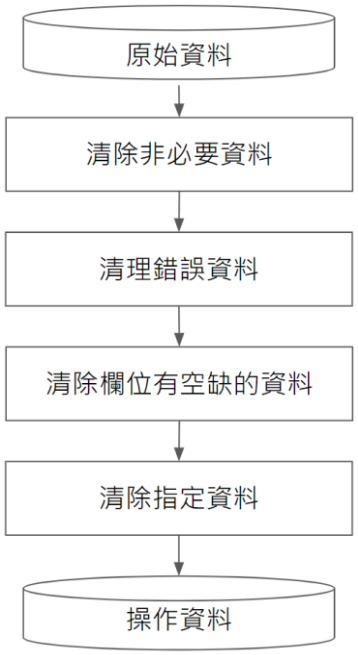


圖 12、實際資料清理流程

3.2.5 資料清理結果

經資料整理後。花蓮資料庫剩餘 1,426,236 筆，私立學校資料庫剩餘 2,148,115 筆，學期與年級分類如表 5：

表 5、操作資料筆數與學期年級範圍

資料庫	花蓮學校	私立學校
學期	109-1 至 110-1 學期	108-1 至 110-1 學期
年級	四 - 九年級	四 - 十二年級
資料筆數	1,426,236	2,148,115

doi:10.6342/NTU202202818

3.3 分群模型建置

在開始建立分群模型前，需要先建立產品端的操作流程，並詳細定義產品描述。再依此定義分群的對象，與分群所需的特徵。本個案的產品端操作流程如圖13。

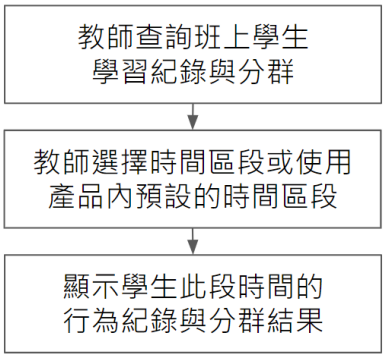


圖 13、產品分群功能操作流程

在此操作流程下，分群模型必須針對某一段時間，將這段時間裡視學生的表現進行分群，因此首先必須決定如何濃縮學生在某段時間內的任務綜合表現，稱為特徵值抽取，我們將在第 3.3.1 節進行詳細描述。

3.3.1 特徵值抽取

由於必須在某段時間內進行分群，所以必須要決定如何濃縮學生這段時間內的綜合任務表現。我們首先將欄位進行分類，分為類別資料與數值資料。類別資料不能進行運算，舉例來說，文本難度並不能平均或加總，因此文本難度是類別資料；而數值資料則可以進行運算，如任務首次正確率，可以平均且具備意義。考慮到欄位內容的多樣性，所以本研究選定數值資料作為特徵值來縮減研究範圍。

本個案經與個案企業產品經理討論，確認分群不應超過 10 種，以簡化教師

doi:10.6342/NTU202202818

使用與參考。若每個特徵值都至少分為兩份，在分群不超過 10 種的前提下，我們最多可以允許三個特徵值。除此之外，特徵值要容易使教師接受，本身必須具備意義且能夠描述學生的任務完成狀況。

在本個案中，我們為了盡可能涵蓋所有任務資料，我們決定不使用文本領域、難度、題型等資料，但若未來能進行更細緻的分群，或許可以採用。我們選用了四個候選特徵值，並以花蓮學校的五年級資料做直方圖分析，此資料共有學生 1774 人，共同擁有 82 個任務，共計 145,468 筆資料，候選特徵值如下：

1. 任務完成率(下文稱之為 Complete Rate)：學生在相同任務中，(資料欄位 MissionStatus = 2 的個數)/總任務數，在此例中總任務數為 82。完成率越高可能可以認為學生較為認真。

2. 平均首次作答正確率(下文稱之為 Correct Rate)：學生在所有任務中，資料欄位所給予的 Correct Rate 平均值。正確率高可能可以認為學生較為認真或閱讀能力較好。

3. 平均首次答題使用時間(下文稱之為 Answer Duration)：學生在所有任務中，資料欄位所給予的 Answer Duration 平均值。在合理時間下，時間越短可能表示學生閱讀能力越佳。

4. 完成任務平均用時(下文稱之為 Working Time)：為每任務資料欄位 FinishTime 減去資料欄位 MissionTime 後取得單一任務的完成任務用時，再將所有任務的完成任務用時平均，取得該值。表示學生得到任務後多快就完成任務。時間短可能可以認為學生較為認真。

我們使用直方圖展示上述四個特徵值候選的分布狀況，如圖 14 至圖 17 所示：

doi:10.6342/NTU202202818

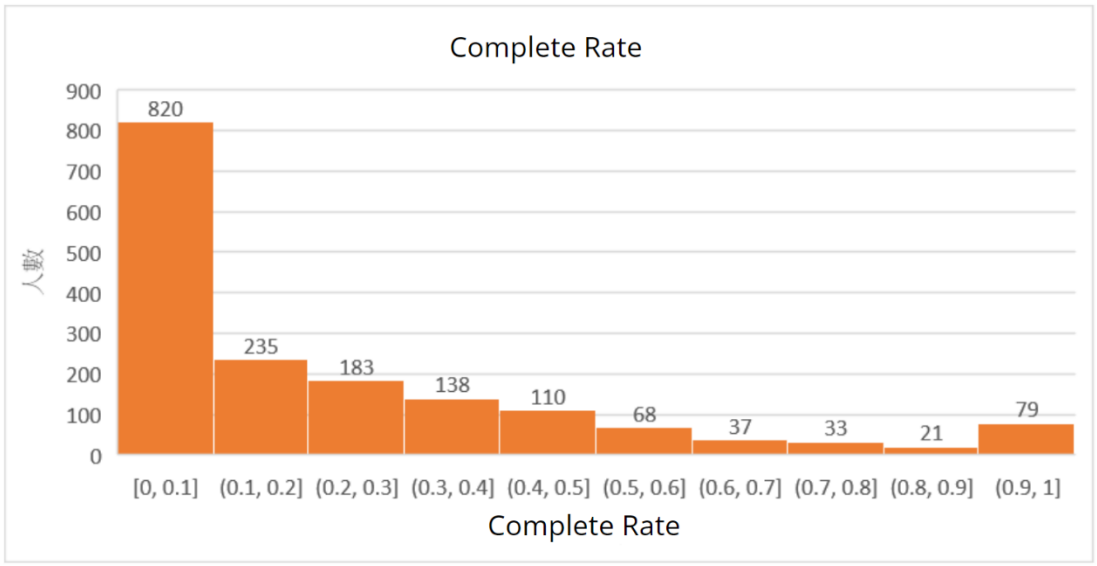


圖 14、Complete Rate 分布狀況

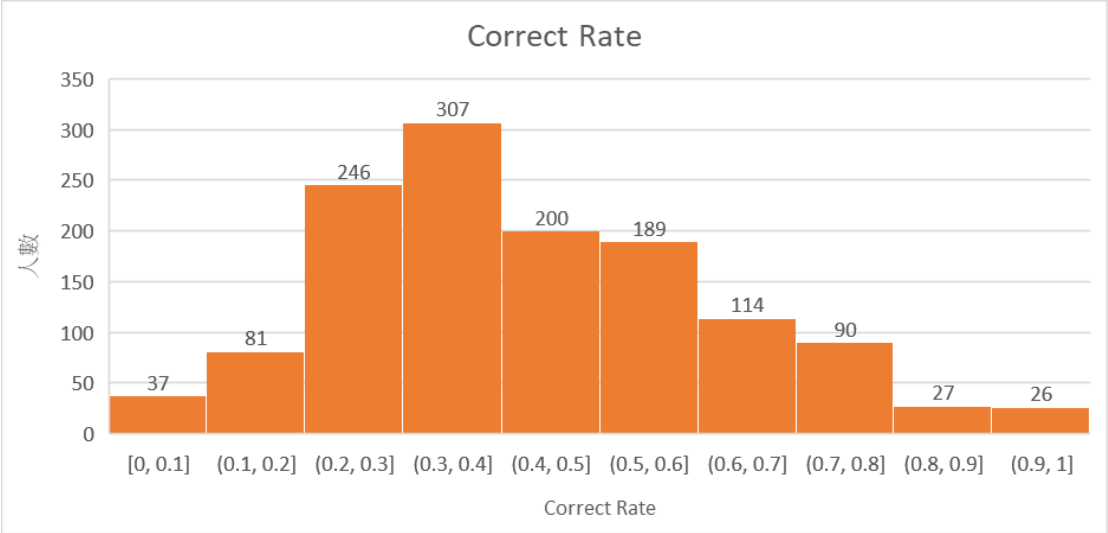


圖 15、Correct Rate 分布狀況

doi:10.6342/NTU202202818

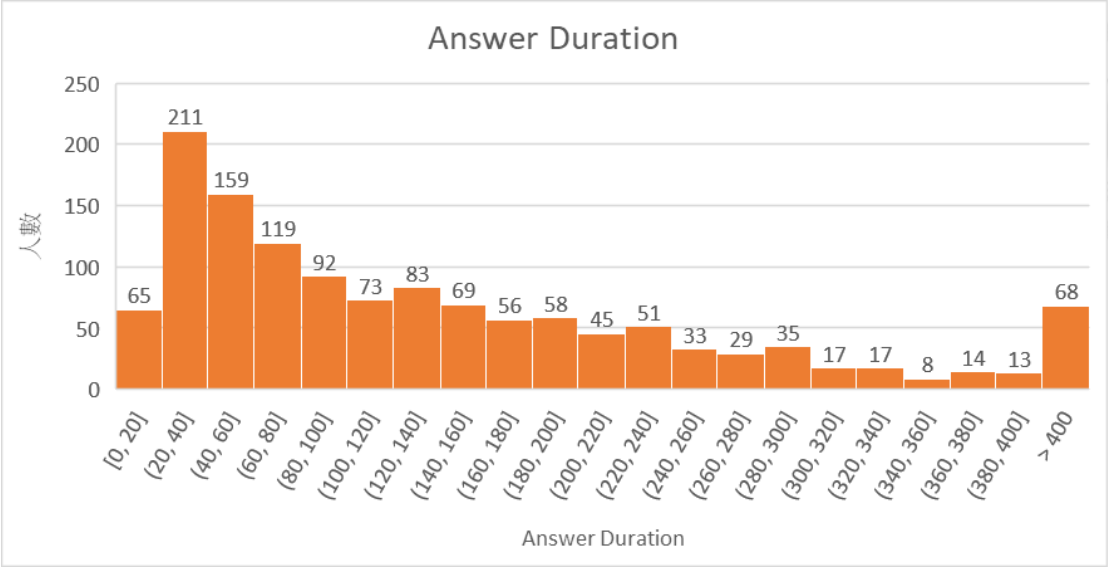


圖 16、Answer Duraion 分布狀況

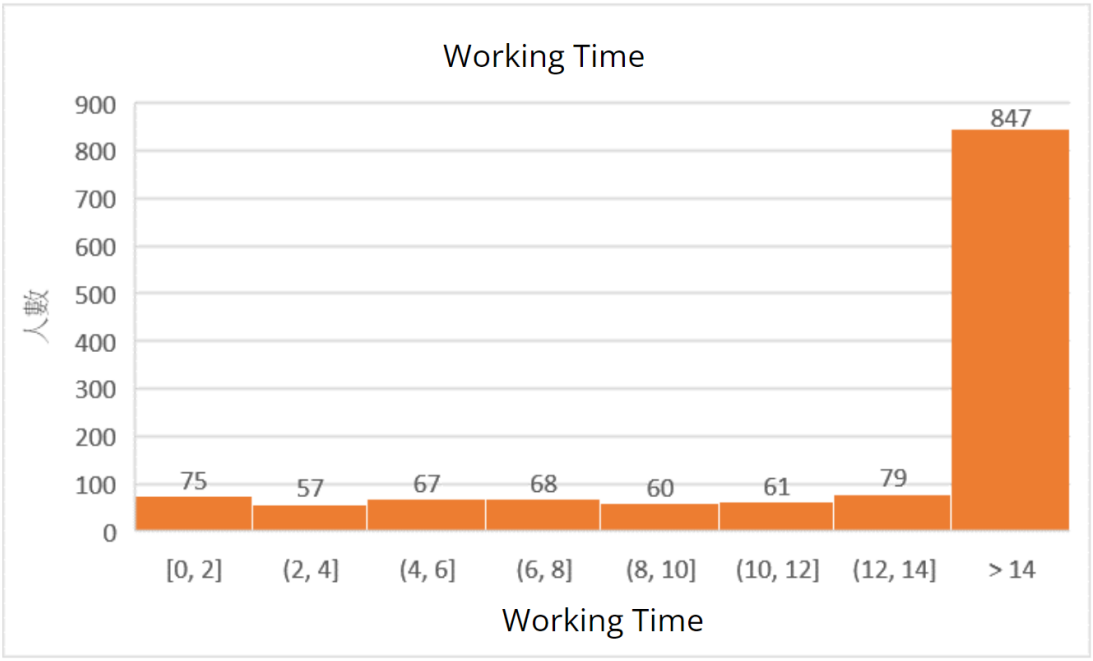


圖 17、Working Time 分布狀況

透過直方圖我們可以發現，完成任務平均用時(Working Time) 絕大多數人都遠超過 14 天，且由於有大量未完成任務，導致有將近一半的學生無法計算此數值，因此最後決定排除此特徵值。所以，特徵值確定為任務完成率、平均任務首次作答正確率、平均首次答題用時。

doi:10.6342/NTU202202818

3.3.2 分群模型方法與驗證

一旦我們選擇了一個時間區段，我們就能將這個時間區段的所有任務資料，依據不同的 userID 區分每位學生，並依上一節的方法製作每位學生在此段時間的特徵值。有了每一位學生的特徵值後，我們至少有兩個候選方法可以建立分群模型，分別為：

1. K-means 分群(K-means Clustering)
2. 以簡單統計量分群

其中 K-means 分群需要指定分群數量，且最後產出為各分群中心點，若運用於本個案，不但需要預先決定分群數量，且在單一特徵值無法明確的區隔，這會造成三個問題：

1. 沒有良好手段能確定最佳分群數量。
2. K-means 分群方法無法在單一特徵值上有明確的區隔，在邊界處會呈現不規則形，以圖 18 為例，我們使用花蓮國小四、五、六年級 110-1 學期特徵值資料，對三個維度進行正規化後進行 K-means 分群，K 設為 8。完成分群後，抽取 Complete Rate 中位數前後各 2.5%總人數共 252 人並於 Answer Duration 與 Correct Rate 兩個維度上繪出。可以發現在邊界處呈現不規則形，這會導致教師難以理解分群的明確意義。
3. K-means 分群以中心點為該分群代表，以此個案而言就是某個學生，由於此個案的分群模型為校際層級，即使教師可以接受分群沒有明確定義，仍難以和教師溝通他校學生所代表意義。

doi:10.6342/NTU202202818

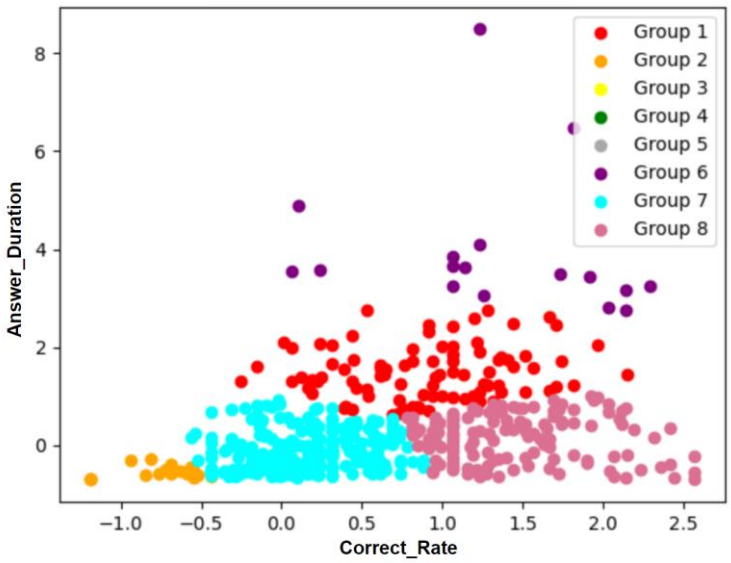


圖 18、K-means 分群結果示意圖

相較於 K-means 分群方法，簡單統計量分群方法不但較容易和教師溝通，也容易依企業需求微調分群模型。因此，本個案在與企業產品經理確認後，使用各特徵值之中位數做簡單統計量分群。為了確定此方案可行，我們選用花蓮國小四五六年級在 110-1 學期的時間區段做驗證。並採用以下分群流程：

1. 去除 Answer Duration 超過 20 分鐘的任務：由於該產品所設計之單次作答時間為五分鐘，因此單次作答時間超過 20 分鐘可能表示該任務的作答有異常發生。
2. 依據上一節的方法製作每個學生的特徵值。
3. 將 Complete Rate 低於 10%的學生分出，視為一群。
4. 其餘剩下的學生計算三項特徵值的中位數。
5. 依計算的中位數將剩下的學生分為八類。

3.3.3 分群結果

doi:10.6342/NTU202202818

圖 19 至圖 21 展示花蓮國小四、五、六年級學生在 110-1 學期的時間區段裡，依前一節流程產生的分群結果。為了表示三維的分群結果，我們針對每個維度的特徵值皆製作兩張二維分布圖，其中每一種顏色為一個分群，線條表示該特徵值之中位數。以圖 19 為例，左圖為 Answer Duration 高於中位數的學生們在 Complete Rate 與 Correct Rate 兩個特徵值的分布狀況；右圖則為 Answer Duration 小於或等於中位數的學生們在 Complete Rate 與 Correct Rate 兩個特徵值的分布狀況。

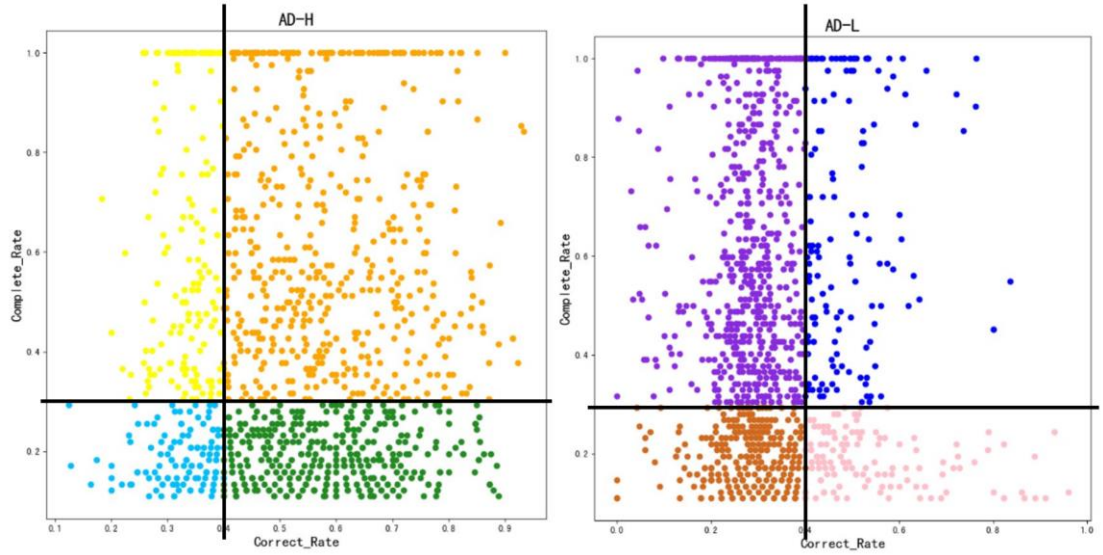
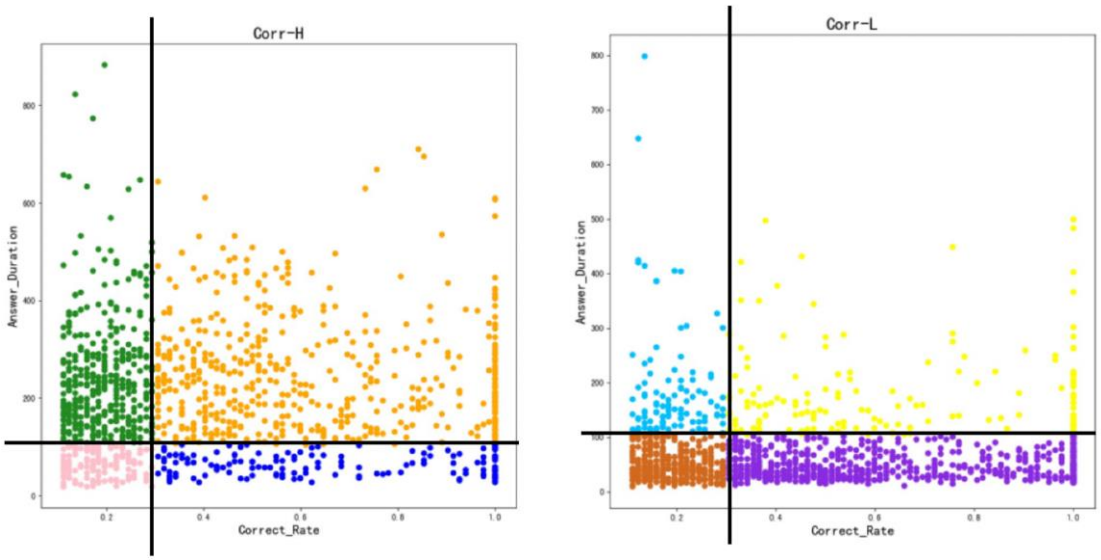


圖 19、(左) Answer Duration 高於中位數；(右) Answer Duration 低於或等於中位數



doi:10.6342/NTU202202818

圖 20、(左) Correct Rate 高於中位數；(右) Correct Rate 低於或等於中位數

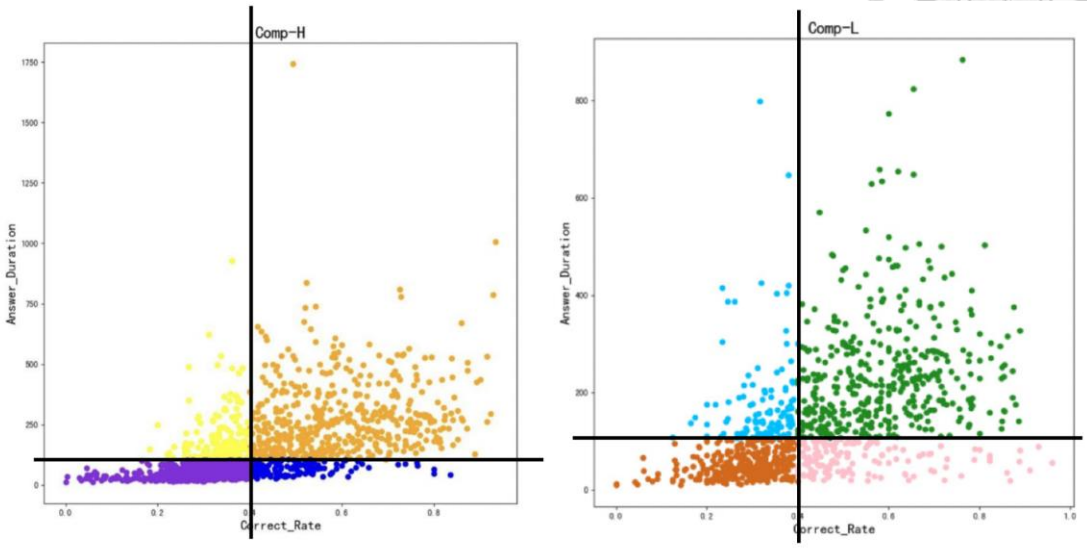


圖 21、(左) Complete Rate 高於中位數；(右) Complete Rate 低於或等於中位數

依上圖所顯示顏色，我們將學生分為 8 群，分別代表以下意義：

1. 分群一，橘色：Answer Duration、Complete Rate、Correct Rate 皆高於中位數，表示此群學生雖然閱讀時間相對較久，但也因此取得較佳的首次作答正確率且較為認真。
2. 分群二，黃色：Answer Duration、Complete Rate 高於中位數，但 Correct Rate 低於中位數，表示此群學生相對認真，但可能閱讀能力需要加強。
3. 分群三，天藍色：除 Answer Duration 高於中位數，其餘皆低於中位數，表示閱讀能力需要加強，且相對較不認真。
4. 分群四，綠色：Answer Duration、Correct Rate 高於中位數，但 Complete Rate 低於中位數，表示學生閱讀時間相對較久，但也取得較佳的首次作正確率，但相比分群一較不認真。

doi:10.6342/NTU202202818

5. 分群五，藍色：Complete Rate 與 Correct Rate 皆高於中位數，且 Answer Duration 低於中位數，表示此群學生閱讀能力優異，能用相較整體學生低的時間取得較好的正確率，而且相對認真。
6. 分群六，紫色：除 Complete Rate 高於中位數，其餘皆低於中位數，表示此群學生雖然看似完成相對多的任務，但實際上有敷衍了事的可能。
7. 分群七，棕色：所有特徵值皆低於中位數，表示此群學生並不太在意此產品提供的服務，相對於分群六更加不認真。
8. 分群八，粉紅色：除 Correct Rate 高於中位數，其餘皆低於中位數，表示此群學生能用相較整體學生低的時間取得較好的正確率，但相對分群五不認真。
9. 分群零：前期 Complete Rate 低於 10%之分群，學生完成任務數量太少，特徵值可能不具意義或是極不認真。

在此分群測試中，共有 5051 名使用者，各分群人數如表 6，可以發現分群零的占比超過 50%，在實務應用時，如果企業端認為此比例過高，可以透過調整分群零的門檻來改變此占比。

表 6、花蓮國小四五六年級學生在 110-1 學期的時間區段各分群占比

分群	名稱	總數	占比(/5051)
0	問題學生型	2595	51.38%
1	縝密思考型	484	9.58%
2	思考緩慢型	96	1.90%
3	危機型	644	12.75%
4	偶而認真型	135	2.67%
5	完美學生型	148	2.93%
6	交差了事型	405	8.02%

doi:10.6342/NTU202202818

7	不在意型	192	3.80%
8	小聰明型	352	6.97%

3.4 實際回饋與企業導入

與個案企業產品經理經過初步討論，認為此分群模型應具備可行性，因此由個案企業產品經理設計並執行實際運用場景測試。測試方法為選擇三個花蓮國小四五六年級的班級，並對該三班的學生進行分群。分群母體為所有花蓮國小的四五六年級學生，在完成分群後將分群結果交付予這三個班級的國文老師。

受測班級的國文教師回饋意見整理如下：

1. 分群結果符合教學現場，與教師平日對學生觀察結果相符合。
2. 可讓教師快速理解學生狀況，不需要逐一檢視學生過往的答題記錄。
3. 分群系統應有助於接手新班級時快速了解學生。
4. 希望可以在初步分群之後再對個別學生顯示其他特徵值來進一步分析學生，如擅長題型或領域等。

其中前三點顯示教師端能理解此分群模型的邏輯，且分群結果對於教師有幫助；第四點則顯示教師端對於分群模型的需求，這也表示建立分群模型對於企業端確實具備其價值。透過此測試，企業端確定此分群模型具備導入價值且具備高度彈性，經成本評估後也確認該分群模型的導入成本不高，因此本模型預計於未來半年進行企業導入。

由第 3.3.2 節的分群流程優化，我們建立了可供軟體開發與一般企業導入分群模型可使用的流程，如圖 22。依循此流程，可以完全複製本個案之分群模型。在本個案中，我們也成功開發此分群模型的軟體系統並交付予個案企業。

doi:10.6342/NTU202202818

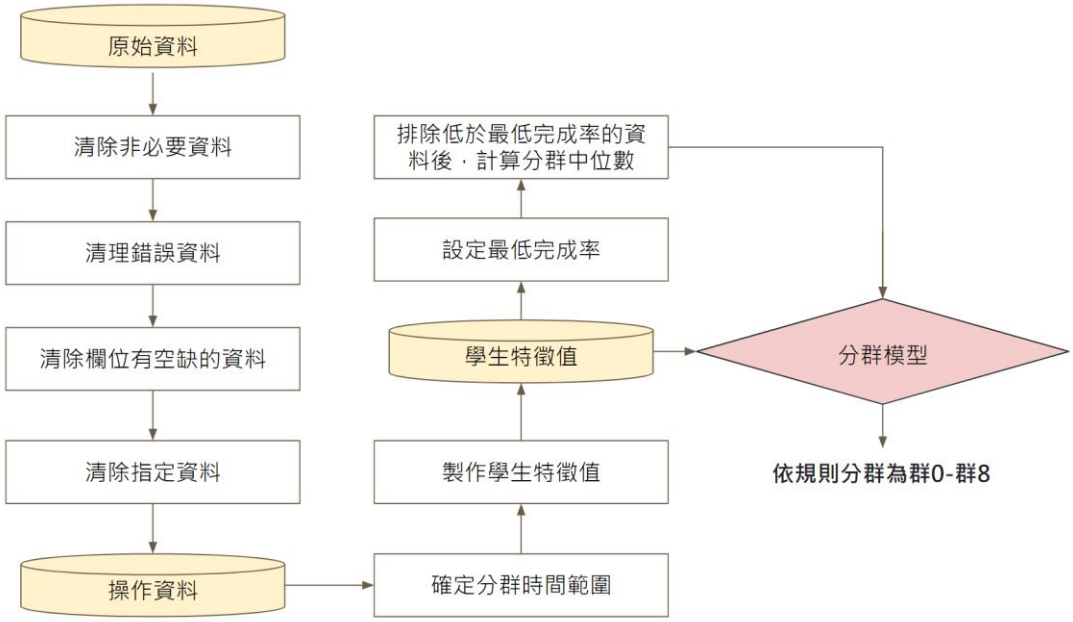


圖 22、個案使用之分群模型完整流程圖

doi:10.6342/NTU202202818

第四章 結論與建議

4.1 研究結論

在本個案研究中，以外部數據分析師的角度，透過與個案企業產品經理共同討論，從確定需求開始，到產品實際功能的測試與落實，並以數據分析的標準流程協助企業端進行系統的導入，我們得到以下結論：

1. 在本案中，對學生的行為資料分析並分群符合教師的需求，並因此提高企業的價值。
2. 依據一般資料分析流程，從敘述性統計開始，逐步進行資料清理、提取特徵值、建立分群模型的流程具備可行性，本研究也揭示可供實作的分群流程圖供參考。
3. 以本個案為例，分群模型在需要考慮分群邏輯溝通與分群模型彈性的情況下，簡單統計分群方法遠較 K-means Clustering 具備可行性。

4.2 後續研究建議

本個案從企業提供的固定資料開始分析，由於大部分欄位只記錄使用者第一次的答題資料，建立分群模型時能使用的特徵值較為有限；但此狀況相當符合企業初次建立數據分析部門的情況，只能取得企業過去所儲存資料。若數據分析部門取得初步成果，數據分析部門應積極與 IT 部門合作，才能蒐集到更多有價值的使用者行為資料。以本個案為例，後續建議研究方向如下：

1. 本個案已初步取得分群功能的預期成果，且依教師所提需求，後續可以針對文本領域、難度、題型等欄位做特徵抽取，如學生「在某段時間內難度為中的任務裡」完成率或「在某段時間內文本領域為世界櫥窗的所有任務中」首次做答平均答對率等，來建立更細緻的分群模型並提供教師做參考。

doi:10.6342/NTU202202818

2. 若資料能涵蓋更多的時間，將能利用該資料建立使用者如何在分群間移動的模型，如馬可夫鏈模型。將模型對應教師所使用的教學方法後，能使個案企業建議教師在不同分群方法應該使用的教學方法，作為下一次產品開發與數據分析的目標。
3. 可以再重新設計使用者行為的蒐集範圍，例如每次作答的答題狀況、上線時數或上線頻率等，可能都可以再抽出分群使用的特徵值。

doi:10.6342/NTU202202818

參考文獻

中文文獻

王郁倫(2021)。台灣企業 60%是數據新手！看好「即服務」商機，戴爾全球喊推 APEX。數位時代。民 111 年 8 月 4 日，取自：https://www.bnext.com.tw/article/65128/dell-as-a-service-data

台灣經濟研究院(2018)。活用數據創造企業新價值。台經社論。民 111 年 8 月 4 日，取自：https://www.tier.org.tw/comment/pec1010.aspx?GUID=82b6c2ce-0afc-4b77-a4c1-4adecd67e1e9

李欣怡(2015)。不懂大數據的 5 大原則、3 大禁忌？小心金礦變災難一場！**數位時代 2015 年 No.251**。臺北市：巨思文化出版。取自：https://www.bnext.com.tw/article/37416/bn-2015-09-17-182050-84

英文文獻

Alsayat, A., & El-Sayed, H. (2016, June). Social media analysis using optimized K-Means clustering. In 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 61-66). IEEE.

Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2008). RFM analysis. In Database marketing (pp. 323-337). Springer, New York, NY.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data

doi:10.6342/NTU202202818

mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 36(3), 4176-4184.

Dogan, Y., Birant, D., & Kut, A. (2013, July). SOM++: integration of self-organizing map and k-means++ algorithms. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 246-259). Springer, Berlin, Heidelberg.

Lin-Siegler, X., Dweck, C. S., & Cohen, G. L. (2016). Instructional interventions that motivate classroom learning. Journal of Educational Psychology, 108(3), 295.

Mojarad, S., Essa, A., Mojarad, S., & Baker, R. S. (2018, June). Data-driven learner profiling based on clustering student behaviors: learning consistency, pace and effort. In International conference on intelligent tutoring systems (pp. 130-139). Springer, Cham.

Patel, V. R., & Mehta, R. G. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. International Journal of Computer Science Issues (IJCSI), 8(5), 331.

Šarić-Grgić, I., Grubišić, A., Šerić, L., & Robinson, T. J. (2020). Student clustering

doi:10.6342/NTU202202818

Based on learning behavior data in the intelligent tutoring system. International Journal of Distance Education Technologies (IJDET), 18(2), 73-89.

Xie, Y., & Phoha, V. V. (2001, October). Web user clustering from access log using belief function. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 202-208).

Zakrzewska, D., & Murlewski, J. (2005, September). Clustering algorithms for bank customer segmentation. In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05) (pp. 197-202). IEEE.

doi:10.6342/NTU202202818