運輸計劃季刊
第五十二卷 第三期
民國一一二年九月
頁 161～頁 190

Transportation Planning Journal
Vo1. 52 No. 3
September 2023
PP. 161～190

多元交通行動服務使用者之套票購買行為分析－以高雄市 MaaS 系統為例

ANALYSIS OF USERS' PURCHASE BEHAVIOR OF MaaS PACKAGES－A CASE STUDY OF THE MaaS IN KAOHSIUNG CITY

盧宗成 Chung-Cheng Lu [1]
簡佑勳 Yu-Shyun Chien [2]
周翔淳 Shiang-Chung Chou [3]
吳東凌 Tung-Ling Wu [4]
陳翔捷 Siang-Jie Chen [5]

(111 年 1 月 4 日收稿，111 年 4 月 7 日第一次修正，111 年 4 月 11 日接受)

摘 要

多元交通行動服務(Mobility as a Service, MaaS)為近年來一項新興概念，國內外許多城市已開始推動 MaaS，了解民眾的套票購買行為及考量因素將有助於主管機關與業者研擬適當的行銷策略以推廣 MaaS。本研究以資料探勘方法建立會員方案續買預測模型，並以高雄市 MaaS 系統會員註冊資料、方案購買記錄及電子票證搭乘記錄為輸入資料。為能改善 MaaS 會員套票購買資料在各方案間之資料不平衡問題，本研究提出以機率分配為基礎之增加抽樣方法(probability distribution-based over-sampling, PDB)，並將此方法與文獻中常用的 SMOTE (synthetic minority over-sampling technique)方法以網路上公開的資料集

1. 國立陽明交通大學運輸與物流管理學系教授 (聯絡地址：10044 臺北市忠孝西路一段 118 號 4 樓，電話：02-23494960，E-mail: jasoncclu@nycu.edu.tw)。
2. 國立陽明交通大學運輸與物流管理學系博士生。
3. 國立陽明交通大學運輸與物流管理學系碩士。
4. 交通部運輸研究所運輸資訊組組長。
5. 交通部運輸研究所運輸資訊組研究員。

進行測試與比較，結果發現 PDB 顯著優於 SMOTE，可進一步應用於處理 MaaS 會員資料，並建構 MaaS 會員方案續買預測模型。經由交叉驗證測試結果發現，經由 PDB 增加抽樣之資料所建構的決策樹及支持向量機模式皆有不錯預測結果，顯示模型具備會員方案續買預測能力。此外，本研究也針對決策樹模型的分支規則進行探討，發現除使用者在各個運具的每月花費會影響續買行為外，月份也是重要的考量因素。

關鍵詞： 多元交通行動服務、資料探勘、決策樹、支持向量機、不平衡資料集

ABSTRACT

Mobility as a service (MaaS) is an emerging concept in recent years, and it has been promoted in many cities around the world. Understanding users' purchase behavior of MaaS packages will help the authorities and operators to promote MaaS. This study develops a package-purchasing prediction model using data mining techniques. The prediction model is built and trained using the data of membership registration, package-purchasing records and iPASS card transit-ridership records of MaaS users in Kaohsiung. Through preliminary data processing and analysis, it is found that significant imbalance exists in MaaS users' package-purchasing records for the different plans, and the imbalance may affect the prediction results of the model. To address this issue, the study proposes an oversampling method, namely probability distribution-based over-sampling (PDB), to generate additional samples. This method is first tested and compared with the SMOTE (synthetic minority over-sampling technique) method commonly used in the literature by using datasets published online, and it is found that the proposed method is significantly better than the SMOTE method. Then the study uses the method to balance the MaaS users' package-purchasing data, and constructs a decision tree model and a support vector machine model for MaaS users' package-purchasing prediction. Through cross-validation test results, it is found that the models constructed by the oversampling data using the simulation method has good prediction results which shows that the prediction model has the ability to predict the users' purchase behavior. This study also discusses the branch variables in the decision tree model, and found that the user's monthly spending on each public transportation mode will affect the package-purchasing of MaaS users. Moreover, month is also an important variable. The results can be used as a reference for MaaS operators to take appropriate actions on marketing based on the results of the users' package-purchasing prediction.

Key Words： *Mobility as a service, Data mining, Decision tree, Support vector machine, Imbalanced dataset*

一、緒　論

多元交通行動服務 (Mobility as a service, MaaS) 的概念為將多元運具整合成為單一運輸服務，透過行動裝置並搭配具經濟效益的選擇性付費方案，提供符合民眾需求的運輸服務。目前許多國家皆積極推動 MaaS，提供民眾多元公共運輸整合服務系統。而我國交通部於民國 105 年辦理「公共運輸行動服務發展應用分析與策略規劃」與「臺北都會區及宜蘭縣交通行動服務建置及經營計畫」，擬定國內引進 MaaS 的適用服務模式、應用範疇、合適場域以及後續推動策略。目前在北臺灣及南臺灣各有一套 MaaS 系統，分別為以雙北、宜蘭為主要計畫範圍之「UMAJI 遊買集」，以及以高雄市為示範場域之「MeN Go 交通行動服務」，本研究則是選定以「MeN Go 交通行動服務」為實證分析對象。

高雄市 MaaS 在 2018 年底正式營運，係以優化使用者體驗、加強行銷推廣、納入更多種類運具服務為推廣策略，並將研擬創新商業模式或合作機制來擴大 MaaS 之應用範圍。目前高雄市 MaaS 已整合捷運、市區公車、公路客運、輕軌、公共自行車、渡輪、計程車、共享電動機車等運具，並透過系統蒐集諸多民眾旅運數據 (如：公車與客運動態、電子票證等)，也因為 MaaS 的推動讓許多原先各別獨立的交通資料得以同時整合，在過去想全面性分析多元交通運具的使用情形多受制於資料量不足且分散，故目前透過導入 MaaS 已大幅度突破原有限制，正是值得深入研究多元公共運輸旅運行為的最佳時機。再則，高雄市 MaaS 自 2018 年正式營運，註冊人數高峰漸已趨緩，故未來如何透過分析旅運資料探討會員購買 MaaS 方案之行為，藉此擬定開發新會員與留住既有會員之行銷策略，將是未來高雄市 MaaS 服務應如何提升營運績效之重要課題 (盧宗成等人 [1])。

近年來，以公共運輸大數據分析為課題之研究中，利用電子票證資料分析使用者旅運行為者如：廖振宇 [2]、鍾智林與李舒媛 [3]、劉芷璇 [4]、林至康等人 [5]、Long and Thill [6]、Morency *et al.* [7]，分別針對使用者轉乘行為、價格彈性、運量時空分佈、旅次起迄特性等為分析課題；而以資料探勘方法探討公共運輸大數據資料之研究中，以悠遊卡乘車資料進行分析者如：林浩瑋 [8]，以手機、穿戴裝置等媒介蒐集旅行軌跡信令資料進行分析者如：Rodríguez *et al.* [9]、Meng *et al.* [10]、王晉元等人 [11]、賴盈臻與邱裕鈞 [12]。在利用 MaaS 營運資料進行分析之研究上，則是以芬蘭之 Whim 較為著名 (Ramboll [13]; Wong *et al.* [14])，而國內則以高雄市之 MeN Go 為濫觴 (盧宗成等人[1])。根據前述分析與文獻回顧內容可得知，目前雖已有諸多研究應用公共運輸大數據資料及資料探勘方法分析使用者之旅運特性或行為，然國內外以 MaaS 所蒐集之整合性旅運數據進行分析者仍屬少數，當中以使用者套票購買方案行為分析者亦無先例，顯示本研究之研究範圍與課題尚具研究空間，值得深入探討。

綜上所述，本研究目的在利用資料探勘方法，分析高雄市 MaaS 系統蒐集的會員註冊資料、方案購買記錄以及電子票證資料，藉此建構 MaaS 方案購買行為之預測模型，分析會員之方案購買行為。其中，考量高雄市 MaaS 會員套票購買紀錄資料具有各方案購買比

例之資料不平衡問題，本研究提出以機率分配為基礎的增加抽樣 (probability distribution-based over-sampling) 方式來改善不平衡資料對於預測模式之影響，避免預測模型發生之過度適配 (over-fitting) 之現象。此外，考量高雄市 MaaS 方案之推廣族群，本研究亦將探討 MaaS 會員在分為一般卡會員及學生卡會員兩類下之預測效果。透過建立 MaaS 方案購買行為之預測模型，模型之預測結果可提供高雄市 MaaS 系統營運策略與規劃之參考，針對即將流失的會員採取適當的行銷策略，藉此提升 MaaS 服務之營運績效。

二、文獻回顧

2.1 公共運輸大數據分析

過去以公共運輸大數據進行分析之研究中，較多是利用電子票證資料分析使用者旅運特性。國內相關研究中，廖振宇 [2] 利用悠遊卡與路線資料庫透過資料探勘和敘述性統計方法，解析 2012 年 10 月之臺北都會區公車轉乘使用者的特性與價格彈性，發現影響旅客的使用因素還是以路線與身分票種等因素為主，而路線因素則發現熱門轉乘地點仍以傳統之交通節點為主，較能滿足旅客需求，便於旅客轉乘。鍾智林與李舒媛 [3] 則以悠遊卡資料進行雙北 Ubike 租借與轉乘捷運的比較分析，利用資料欄位串聯與羅吉斯迴歸模式判斷 Ubike 租借與捷運進站之時間關係是否有轉乘並進一步探討其行為偏好等，研究發現大多數每月租借不到 2 次的使用者是於假日或平日昏峰以休閒遊憩等目的為主，每月租借超過 2 次的使用者則是以平日通勤為主，轉乘比例較高。劉芷璇 [4] 以臺中市公車為例，利用公車電子票證資料分析公車與乘客旅次資訊，找出票種、業者與路線 (運量、旅次長度、起迄點) 特性，利用 VB 之 ADO 資料庫處理程式整理資料、並使用 MATLAB 繪出車上旅客人數變化情形，作為公車業者排班與調度之依據。林至康等人 [5] 利用市區公車電子票證一段次刷卡紀錄，並結合電子票證與地區土地分區之資料，建立一段式電子票證刷卡資料的 3 階段旅次訖點推估演算法，藉此了解使用者於大眾運輸之旅次起迄特性。國外相關研究上，Long and Thill [6] 利用公車電子票證資料、家戶旅次調查資料及地圖等，找出使用者工作地及家戶地點，同時找出其通勤旅次路線，分析通勤旅次的模式。Morency *et al.* [7] 以智慧票證資料為基礎，分析民眾轉乘公共運輸之特性，並探討不同類型使用者間之差異。

隨著資料探勘方法成為近年顯學，亦有不少關於資料探勘方法應用在交通運輸之文獻。林浩瑋 [8] 以悠遊卡乘車資料，進行淡水的捷運與公車兩運輸系統的旅客通勤行為比較分析，該分析先利用群聚分析法與資料分群、關聯法則等研究方法進行比對，研究發現捷運通勤族與公車通勤族之通勤行為有明顯區別。Rodríguez *et al.* [9] 藉由穿戴裝置及搭配手機應用程式完整蒐集旅客的移動數據，並加以整理分析出各旅客族群特性與需求，利用蒐集到的地理定位資訊、時空資料進行集群分析，並從結果顯示該地區旅客可分為兩個主要集群、四個子集群，共可將不同旅客特性細分為五個種類特性，期望可以結合更多面向之資訊 (家戶資料、醫療保健等)，使該地區之公共運輸更滿足使用者需求。Meng *et al.* [10] 根

據旅行軌跡資料、POI (Point of Information) 資料及社交媒體資料來推論個別旅次目的，並運用動態貝氏網路 (Dynamic Bayesian Network, DBN) 模型取得重要因子；研究中除使用異質性資料外，亦利用 POI 的熱門程度來推斷旅次目的，藉此從 DBN 分析結果得知旅次鏈順序關係；研究結果則顯示以前述資料作為建模基礎可獲得不錯的準確率。王晉元等人 [11] 利用資料探勘方法，分析觀光地區之行動信令資料，藉此比較潛在公共運輸旅客的觀光旅運需求樣態以找出服務缺口，並提出改善建議。賴盈臻與邱裕鈞 [12] 同樣利用行動信令資料，採用基因模糊邏輯控制與 K 近鄰、決策樹及隨機森林等資料探勘方法，探討使用者之運具選擇行為。

2.2 MaaS 營運資料分析

由於 MaaS 系統係透過整合多元運具至單一平台，並利用行動裝置 App 提供民眾一站式消費 (one-stop-shop) 之公共運輸服務，故 MaaS 營運業者亦可藉此蒐集使用者於各運具之使用資料，藉此進行應用大數據方法針對民眾運具使用特性進行分析，並提供政府單位作為研擬公共運輸改善策略之參考。

目前實際營運之 MaaS 案例仍相對較少，而利用 MaaS 系統之營運資料進行分析者更是少數，透過 MaaS 營運資料進行分析之研究中，係以 Ramboll [13] 較為著名，芬蘭 MaaS 服務 Whim 之營運商 MaaS Global 於 2019 年委託顧問公司－Ramboll 分析其 2018 年 1 月至 12 月之營運數據，並與赫爾辛基當地未使用 MaaS 服務居民之旅運資料做比較，Ramboll [13] 當中分析之主要項目包含：使用者特性分析、運具市占率分析、複合運具 (轉乘) 使用情況分析、旅次長度 (距離、時間) 分析、旅次量分析等，研究中除發現 Whim 用戶使用公共運輸之比例較高外，亦發現 Whim 可透過副大眾運輸工具解決使用者第一哩及最後一哩路之問題，而前述分析資料皆將作為業者後續研提行銷策略、改善使用者公共運輸服務體驗之參考。

Wong *et al.* [14] 則是進一步以 Whim 之使用者數據進行分析，並發現 MaaS 系統對使用者體驗有良好之正向效果、對社會整體使用效率及降低碳排放皆有一定效果，且 MaaS 系統以使用者導向之服務模式，改變了以往傳統較為供給者導向之運輸服務。盧宗成等人 [1] 則是以高雄市之 MaaS 服務－MeN Go 為對象，分析會員之共享運具使用特性分析，藉此針對共享電動機車使用者特性、影響共享電動機車使用之因素、公共運輸與共享電動機車轉乘熱點及效益、公共運輸與共享電動機車轉乘空間縫隙等進行分析，並研提各項分析之改善建議。

2.3 資料不平衡問題

現實情況中具有許多資料不平衡之問題，例如：醫療診治、顧客流失率、風險管理等，這些情況下的少數類別通常不能被忽視，倘若於資料分析過程中判斷錯誤便可能使結果產生極大的偏誤，Barua *et al.* [15] 便有提到在資料不平衡的資料集裡如何判斷出少數類別是不

容易的事情。目前處理不平衡資料常使用抽樣方法，藉由改變訓練資料的分布來降低、消除資料的不平衡現象，而 under-sampling 與 over-sampling 為抽樣方法的基本呈現，在處理較複雜資料的情況下，over-sampling 的表現會比 under-sampling 好。而在眾多 over-sampling 的方法中，SMOTE (synthetic minority over-sampling technique) 演算法是最常被用於處理不平衡資料問題的一種，SMOTE 是由 Chawla *et al*. [16] 提出，主要是透過 *k* 鄰近法 (*k*-nearest neighbors)，於少數類別中隨機選擇一個樣本點，並找出他的 *k* 個鄰近點，以人造合成方式產生倍數化的少數類別資料來改善少數類別資料於分類上的偏誤。以 SMOTE 方法改善資料不平衡之研究繁多，陳威廷 [17] 以汽車租賃信用風險評估為課題，並利用 SMOTE 演算法處理高風險與低風險資料不平衡的問題，並透過應用二階段邏輯斯迴歸來建構模型，研究結果亦顯示此模型可有效用於客觀評估汽車租賃申請案件的違約風險；王銘亨 [18] 則以臺南市之閃光號誌路口交通事故特性為課題，並以 SMOTE 方法改善事故資料中死亡事故與受傷事故之資料不平衡問題。

SMOTE 演算法雖在 over-sampling 上有不錯的表現，且具有方便操作、運算快速等優點，惟 SMOTE 演算法係以 *k* 鄰近法作為資料產生基礎，故 over-sampling 過程中亦可能產生不合理之人造資料點，進而使原始資料之型態改變。也因此，除較常見之 SMOTE 方法外，過去亦有考量原始資料特性或分布情況以解決資料不平衡之研究。Chen *et al*. [19] 則提出以資訊粒為基礎的資料探勘方法 (information granulation based data mining approach)，該方法係透過計算資料集在各個類別樣本 (class label) 之資訊粒 (information granule) 以設定樣本權重，藉此解決資料不平衡問題並提升資料探勘方法之預測準確率，而透過使用不同資料集來進行驗證後，也證實其所提出之方法可以有效解決資料不平衡問題。陳世承 [20] 提出少數類別抽樣技術之改良方法來解決資料不平衡的問題，此方法同時衡量少數類別和多數類別之密度分布，並以此作為權重衡量的基礎，可減少產生不合理之人造資料而使分類效果不佳的情形發生。

2.4 文獻評析

綜整上述文獻回顧內容，目前雖已有不少研究以公共運輸大數據分析為研究課題，而運用之資料集中亦不乏以電子票證、手機信令資料、公共運輸班次資訊等，藉此分析使用者之起迄分布、運具選擇、轉乘行為等旅運特性，然目前以 MaaS 所蒐集之整合性旅運資料進行分析者仍屬少數，而以 MaaS 服務中之使用者套票方案購買行為為研究課題者則尚為闕如，顯示本研究以高雄市 MaaS 之旅運資料建構預測模型，並以 MaaS 之套票購買行為為研究課題，除在學術面尚具研究空間外，於實務面應具應用價值。此外，為能避免資料不平衡問題而使分析結果產生偏誤，亦有諸多研究建議採以 over-sampling 方法進行改善，當中又以 SMOTE 演算法為較常應用於解決資料不平衡問題之 over-sampling 方法，然考量 SMOTE 演算法可能會使資料型態發生改變之缺點，故本研究亦將嘗試就原始資料之分佈特性著手，提出不同於 SMOTE 演算法之 over-sampling 方法以改善其缺點，藉此提升本研究於學術面之貢獻性。

三、研究方法

3.1 問題描述

高雄市 MaaS 系統自 2018 年底正式營運，至 2021 年底已經滿三年，總會員數已經超過 30,000 人，除了在國內 COVID19 疫情較為嚴重的 2021 年 5 月到 9 月這段期間外，每月購買高雄 MaaS 系統套票的會員人數約介於 1500 至 2500 人之間，顯示高雄市 MaaS 系統已具有一定的使用量，惟每月固定購買套票人數仍與總會員人數具有相當差異，顯示 MaaS 服務之推廣仍具有長足之改善空間，加以考量目前有關國內 MaaS 的研究較少，無法清楚明白每個月使用 MaaS 系統的這些會員在方案選購上的特性與規則。為能瞭解 MaaS 會員使用者的套票方案購買行為，本研究透過交通部運輸研究所計畫經費補助，使用高雄市 MaaS 系統所提供之 2018 年 11 月至 2019 年 5 月，共 7 個月份之電子票證資料與 MaaS 系統會員相關基本社經資料[6]，來分析在高雄市 MaaS 系統這段時間的營運下，MaaS 系統會員的方案套票購買行為，從目前現有的 MaaS 系統會員資料使用資料探勘方法來尋找哪些因素可能為這些會員續買方案或是不續買方案的原則，並且建立 MaaS 系統會員的套票方案續買預測模型。值得一提的是本研究僅探討正常情形下高雄 MaaS 系統會員之套票購買行為，故 COVID19 疫情對 MaaS 系統使用行為之影響不在本研究的範圍內；根據高雄 MaaS 系統業者訪談結果，在 2021 年 9 月疫情趨緩後，MaaS 系統之每月使用量已經逐漸恢復到疫情發生前的水準。

本研究根據高雄市 MaaS 系統所提供的會員方案購買紀錄資料發現，高雄市 MaaS 系統中每月使用 MaaS 服務之會員人數約在 5000 人左右，若以 2019 年 4 月之 MaaS 方案購買數量為例 (如表 1)，高雄市 MaaS 系統會員在各種方案購買或未續買的比例上有著明顯的資料不平衡現象，因此很有可能因為資料本身的因素導致建立起來的套票方案續買預測模型會失準，因此本研究會在建立預測模型前，先處理資料不平衡的問題，再以處理後的資料建立預測模型。

表 1　2019 年 4 月 MaaS 方案購買數量情形

所購買之方案	購買人數
公車暢遊方案	166
公車+客運暢遊方案	91
無限暢遊方案	2096
未續買方案	2378

註：未續買方案者為同一使用者相較於 2019 年 3 月未持續購買方案

[6] 各項分析資料皆已由 MaaS 營運業者進行個人資料去識別化作業後再提供予本研究進行分析使用，故各筆資料皆不會與特定個人有連結。

本研究方法流程分為三階段，第一階段 (準備階段) 利用網路公開資料進行不平衡問題解決方法之篩選，第二階段 (訓練階段) 則利用電子票證資料進行資料探勘模型的分析與建置，第三階段 (預測階段) 則利用資料探勘模型進行會員套票方案購買之預測，主要資料分析流程如圖 1 所示。

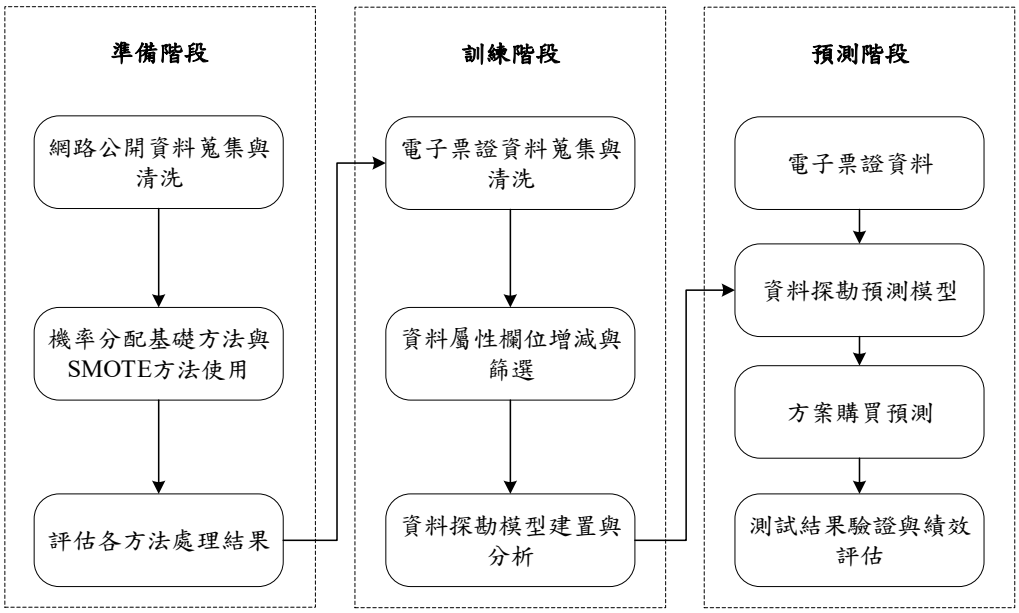


圖 1　資料分析流程圖

3.2 機率分配基礎增加抽樣法

為了減少 MaaS 會員方案購買預測模型受到前述資料不平衡問題影響，除利用過去研究中常採用的 SMOTE 方法對少數類別樣本進行 over-sampling 外，考量 SMOTE 方法係以現有資料集為基礎，並以 k 鄰近法來合成產生倍數化資料達到 over-sampling 的目的，因此在隨機抽取資料點以合成新資料點時，將可能造成資料型態改變；此外，若原始資料之分布情況較為不均或分散時，人工產生之新資料點亦可能會圍繞在較分散的資料點周邊，進而加劇資料點的分散情況，並使 over-sampling 後的樣本資料更難被模式分類。有鑑於以 SMOTE 進行 over-sampling 時可能導致之問題，本研究係以資料原始之分佈型態為出發點，並考量不同樣本資料皆有其特定之機率分配型態，若能以適合之機率分配作為少數類別樣本進行 over-sampling 之基礎，亦可使 over-sampling 結果更符合原始資料型態，並提升分析結果的準確性。

基此，本研究提出以機率分配為基礎之增加抽樣 (probability distribution-based over-samplin, PDB) 方法來增加少數類別樣本，藉此處理 MaaS 會員方案購買資料之不平衡問題。為能確認本研究所使用的 PDB 方法可以有效處理資料不平衡問題，本研究先使用

Malerba [21] 於 UCI 網站所公開之分析資料集－Page Blocks Classification Data Set (PBC Data Set) 為例來說明 PDB 之執行方法，其執行流程如圖 2 所示。

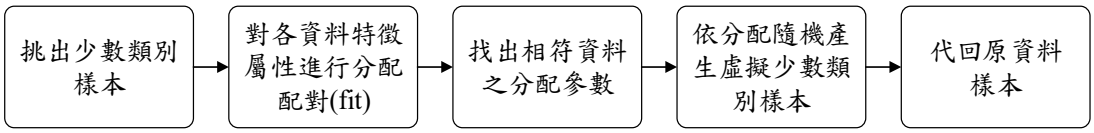


圖 2　PDB 方法處理資料不平衡問題之執行流程圖

1. 挑出少數類別樣本

使用 PDB 方法時，首先需觀察資料集中的多數與少數類別資料樣本，並取出其中的少數類別資料樣本進行 over-sampling，增加少數類別資料的數量至與多數類別資料數目相近。以 PBC Data Set 為例，該資料集中共有 5 個類別 (class label) 之資料，其數量與比率如表 2 所示。為能簡化說明範例，本研究係將 4 類較少數類別資料整合為同一類別，僅留下多數與少數等兩大類別之樣本資料；由彙整表中可發現，多數類別樣本與少數類別樣本之數量確具有明顯差異。

表 2　Page Blocks Classification 範例資料集各類別樣本資料分布情況

樣本資料類別		數　量		百分比 (%)	
多數類別樣本(1)	text	4913		89.8	
少數類別樣本(0)	horiz. line	329	560	6.0	11.2
	picture	115		2.1	
	vert. line	88		1.6	
	graphic	28		0.5	
總　計		5473		100	

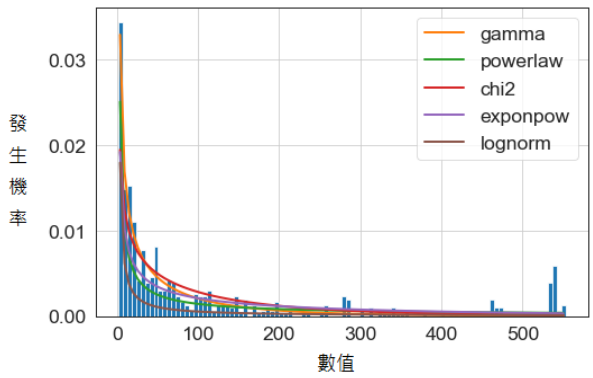
資料來源：Malerba [21]；PBC Data Set。

2. 對各資料特徵屬性進行機率分配配適

經識別資料集中的少數類別資料後，其次需針對少數類別資料的各項特徵屬性進行機率分配的配適。在連續型特徵屬性上，本研究係使用 python 的 fitter 函式庫進行適配，fitter 函式庫提供了 80 種連續機率分配函數提供使用者進行使用，因此本研究將少數類別資料的連續型特徵屬性－Length，與 80 種連續機率分配函數進行配適，並且以配適後誤差平方和 (sum square error) 最小之分配函數作為下一階段 over-sampling 之參考，其契合對比示意圖如圖 3 所示，由彙整結果可得知 gamma distribution 具有最小之誤差平方和，為最佳之配

適函數。至於離散型特徵屬性則需先對特徵屬性之種類進行彙整，再依各個種類之數量與比例進行統計以瞭解其機率分配狀況，藉此作為下一階段 over-sampling 之參考基礎；由於 PBC Data Set 中並無離散型特徵屬性，故本研究以 PBC Data Set 中之連續型變數-Height 為例，透過將連續型變數離散化，藉此繪製圖 4 之離散型特徵屬性趨勢統計示意圖；倘若某一離散型特徵屬性只有 3 種數值，分別為：1、2 與 3，且此 3 種數值之比例又分別約為 95%、3.75%與 1.25%，則後續便將以此比例為參考值對離散型特徵屬性進行 over-sampling。

分配函數	誤差平方和
gamma	0.000134
powerlaw	0.000343
chi2	0.000390
exponpow	0.000449
lognorm	0.000784



資料來源：Malerba [21]；以 PBC Data Set 中之連續型特徵屬性－Length 為例進行繪製

圖 3　最佳前 5 機率分配與連續型特徵屬性契合比對示意圖

Height 部份資料範例	
連續型特徵屬性	離散化後特徵屬性
95	1
98	1
100	1
101	2
105	2

連續型特徵屬性範圍	離散化後特徵屬性	次數	百分比
1~100	1	532	95.00%
101~200	2	21	3.75%
200 以上	3	7	1.25%
加　總		560	100%

資料來源：Malerba [21]；以 PBC Data Set 中之連續型特徵屬性－Height 為例進行繪製

圖 4　離散型特徵屬性趨勢統計示意圖

3. 找出相符資料之機率分配參數

在找出與少數樣本特徵屬性最契合的機率分配後，為能利用此分配函數進行虛擬資料樣本生成，本研究同樣使用 fitter 函式庫，將特徵屬性與該機率分配進行配適，藉此計算式(1)所列 gamma distribution 機率密度函數對應之各項參數；其中，gamma distribution 可

表示為Γ(α)，*式(1)*中α將影響函數陡峭程度，β則會影響散佈程度，各項參數之校估結果如圖 5 所示。

$$f(x) = x^{(\alpha-1)} exp^{-\frac{x}{\beta}} / \Gamma(\alpha)\beta^{\alpha} \quad (1)$$

α (陡峭程度)	β (散佈程度)
0.43319	128.35024

圖 5　Gamma distribution 各項參數計算結果

4. 依機率分配隨機產生虛擬少數類別樣本

在得出分配的參數後，便使用該分配並代入其對應的參數值來依分配隨機生產虛擬少數類別樣本，而生成後的樣本總數將以多數類別樣本數為基礎進行設定，藉此令兩者之樣本數量相近，其範例如表 3 所示。

表 3　樣本數於 over-sampling 前後之差異

類　別	原始樣本數量	更新後樣本數量
多數類別樣本(1)	4913	4913
少數類別樣本(0)	560	4913

註：Over-sampling 是以 Malerba [21]公開於 UCI 網站之資料集-PBC Data Set 為例

5. 代回原樣本資料

將生產出來的虛擬少數類別樣本代回原始樣本資料成為一個新樣本，之後便可依照使用者的分析需要再做調整以進行不同方法的的資料探勘分析，而本研究將使用 PDB 方法產生的新樣本與 SMOTE 產生的樣本進行比較。

3.3 績效指標

在兩個類別的資料不平衡問題評量上，通常數量較少的類別會被稱為 positive class，而數量較多的類別則被稱為 negative class。confusion matrix 是很典型的評估方法，其範例如表 4 所示；列代表真實的類別標籤，行代表資料探勘方法所預測的類別標籤，TP (True Positive) 是指被資料探勘方法分類正確的少數類別資料個數，FN (False Negative) 則是指被資料探勘方法分類錯誤的少數類別資料個數，而 FP (False Positive) 則是指被資料探勘方法分類錯誤的多數類別資料數量，TN (True Negative) 是指被資料探勘方法分類正確的多數類別資料數量，其中從 confusion matrix 衍生出的常用指標詳述如下：

表 4　confusion matrix

預測類別 / 實際類別	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

1. Recall (= TPrate = $\frac{TP}{TP+FN} = \frac{TN}{FP+TN}$)

recall 是指在所有 positive class (或 negative class) 中，資料探勘方法分類正確的比例。

2. Precision (= $\frac{TP}{TP+FP} = \frac{TN}{FN+TN}$)

precision 是指在所有被資料探勘方法分類為 positive class (或 negative class) 的資料中，資料探勘方法分類正確的比例，與 recall 之間衡量的方式在於資料探勘方法在 positive class (或 negative class) 中被分類錯誤的部分所造成的代價是否很大，如果是的話則建議選擇看 recall。

3. F-measure (= $\frac{(1+\beta^2)*recall*precision}{\beta^2*recall+precision}$)

F-measure 的參數 β 是可以自行調整的，用以調整 recall 與 precision 之間的權重，不過通常情況下會被設為 1，而此時的 F-measure 也就是常見的 F1-score，而當 recall 與 precision 都很高時則 F-measure 也會跟著高，故 F-measure 也是可以被拿來作為評估資料探勘方法的分類能力指標。

3.4 Over-sampling 方法測試與比較

為能比較 PDB 及 SMOTE 方法於 over-sampling 結果之優劣，本研究共使用 15 個 UCI 網路提供之公開資料集為測試資料 (UCI [22])，在經過 PDB 方法以及 SMOTE 演算法 over-sampling 後，分別將每個資料集經 over-sampling 後之新樣本套入 decision tree、kNN 與 logistic regression 等方法，並且依照 3.3 節所提及的各項指標來進行比較，藉此驗證本研究所提出之 PDB 方法，其 over-sampling 成效可與過去常用之 SMOTE 演算法比擬，並能進一步應用於處理本研究所使用的 MaaS 會員交易資料。

表 5 為前述 15 個網路公開資料集在使用 PDB 方法以及 SMOTE 演算法進行少數類別的 over-sampling 後，分別使用 Decision Tree、KNN、Logistic Regression 三種資料探勘方法比較 Recall、Precision、F1-score 的勝負次數統計，可以看出無論是個別的任一資料探勘方法還是整體加總 (total count) 皆是 PDB 方法的勝利次數高於 SMOTE。

表 5　PDB 方法與 SMOTE 演算法之勝負比

項目	Decision Tree			kNN			Logistic Regression			Total Count		
	PDB	SMOTE	平手	PDB	SMOTE	平手	PDB	SMOTE	平手	PDB	SMOTE	平手
Recall	9	4	2	10	4	1	8	5	2	27	13	5
Precision	8	5	2	11	2	2	9	5	1	28	12	5
F1-score	8	4	3	10	4	1	9	4	2	27	12	6

資料來源：UCI [22]

而為了明確瞭解 2 種方法是否在分類的表現上有顯著的差異，本研究除了統整出表 4 外也額外再使用無母數統計檢定中的雙樣本中位數差異檢定 (Wilcoxon signed rank test) 去比較 PDB 與 SMOTE 等兩種方法。首先，本研究計算每一個網路公開資料集在 PDB 方法與 SMOTE 之間的準確度差異，求出差異v_i (i=1,2,3,⋯,15；差異是絕對值)，接著由小到大排序所有v_i，並且依據此排序給予每一個v_i一個排名分數，最小的v_i給予排名分數 1，第二小的v_i給予排名分數 2，以此類推；而當發生數個v_i的差異是相同的情況時，則將v_i的排名分數平均，再接著將每一個v_i依據其原始差異是屬於正還是負進行分類並分別加總，正分類加總分數以R^+表示，負分類加總分數則以 R^-表示，R^+和R^-之中的最小值則可以被轉換成 p-value；綜上，本研究之假設檢定如下：

H_0：PDB 方法下之分類表現沒有明顯優於 SMOTE 演算法下之分類表現

H_1：PDB 方法下之分類表現明顯優於 SMOTE 演算法下之分類表現

表 6 為使用 PDB 方法以及 SMOTE 演算法進行少數類別 over-sampling 後，分別在 decision tree、kNN、logistic regression 三種資料探勘方法下，對於上述假設檢定所計算出 Recall、Precision、F1-score 之 p-value；由彙整表中可以發現，所有的 p-value 皆於顯著水準 α = 0.1 時具顯著差異，故可拒絕虛無假設 H_0，得出在 PDB 方法下之分類表現顯著優於 SMOTE 演算法下之分類表現的分析結果。

表 6　比較 PDB 方法與 SMOTE 演算法之假設檢定值 p-value 與結果

項　目	p-value			Count
	Decision Tree	kNN	Logistic Regression	
Recall	0.095 *	0.081*	0.095 *	3
Precision	0.095 *	0.046**	0.085 *	3
F1-score	0.094 *	0.085*	0.093 *	3

註：***p<0.01, **p<0.05, *p<0.1

根據上述 3 種資料探勘方法測試後，可發現本研究欲使用之 PDB 方法表現能夠優於現今常被使用的 SMOTE，PDB 方法所產生的虛擬少數樣本資料能夠更貼近原始樣本資料的樣貌，故本研究除以 SMOTE 方法外，亦將多考量採用 PDB 方法來處理 MaaS 電子票證資料的資料不平衡問題，接著再使用資料探勘方法建立預測模型，藉此分析 MaaS 會員的方案購買行為。

四、資料說明與處理

4.1 研究資料說明

為建立高雄 MaaS 系統會員之方案購買預測模型，本研究擬納入之資料包含：系統會員註冊資料、方案購買紀錄、電子票證搭乘紀錄等三項，各項資料皆已由 MaaS 營運業者進行個人資料去識別化作業後再提供予本研究進行分析使用，故各筆資料皆不會與特定個人有連結，其說明如下：

1. 會員註冊資料

本研究使用之資料為高雄 MaaS 系統從試營運開始截至 2019 年 5 月 (包含 5 月) 之所有會員的註冊相關資料，該資料包含：識別碼、註冊日期、會員出生日期、縣市、鄉鎮區、卡片類型、性別等，共計 7 個欄位資料，總資料筆數共 17,789 筆。

2. 方案購買紀錄

本研究使用之資料為高雄 MaaS 系統於 2018 年 11 月至 2019 年 5 月共計 7 個月份，此期間內所蒐集之會員方案購買資料，包含：識別碼、卡片類型、所購方案、方案費用、方案有效持續時間、購買方案的下單日期等，共計 6 個欄位資料，總資料筆數共 742,568 筆。

3. 電子票證搭乘紀錄

與項目 2 相同，本研究同樣使用 2018 年 11 月至 2019 年 5 月，共計 7 個月份之會員票卡交易紀錄資料，包括：識別碼、運具別、業者別、上車時間、下車時間、上車站代碼、下車站代碼、扣款金額等，合計 8 個欄位資料，總資料筆數共 1,048,576 筆。

4.2 資料處理

針對 4.1 節所述之分析資料，本研究為能夠順利進行資料探勘方法分析與套票方案購買預測模型建立，將利用 python 程式語言針對原始資料進行處理，其流程圖如圖 6 所示。

圖 6　資料分析流程圖

1. 欄位處理

本研究首先針對各項原始資料之欄位進行篩選及換算，以利後續進行建模；其中，高雄 MaaS 系統會員註冊資料的出生日期將換算成會員年齡；高雄 MaaS 系統會員方案購買紀錄資料的主要功能為記錄每一位會員在特定時間點的方案持有與否以及持有何種方案，故本研究將高雄 MaaS 系統會員註冊資料中「卡片類型」之重複欄位，以及高雄 MaaS 系統會員方案購買紀錄資料主要功能中較無關聯性之「方案費用」、「購買方案的下單日期」兩欄皆剔除；高雄 MaaS 系統會員電子票證搭乘紀錄資料因「上車站代碼」與「下車站代碼」缺乏相對應的 API 資料，且「下車時間」一欄的紀錄上有大量的空缺，故剔除上述三個欄位。

2. 資料勾稽

由於高雄 MaaS 系統會員方案購買紀錄資料重複性過高，同一位會員購買單一方案即會產生多筆資料，而其中卻僅有「方案有效持續時間」一欄位有變動，考量本研究後續將以「月」為單位進行資料探勘方法分析與建立套票方案續買預測模型，故需進行資料勾稽整理高雄 MaaS 系統會員的方案購買紀錄；勾稽方式上，本研究首先以高雄 MaaS 系統會員註冊資料的識別碼欄位作為資料欄位串聯基礎，藉此將高雄 MaaS 系統會員方案購買紀錄的每筆資料對應至相符的會員識別碼與日期，並紀錄當下所持有的方案狀態，進而將方案持有情況由文字轉為數值化如表 7，資料勾稽展生之新資料庫範例則如圖 7，勾稽後可供分析之總資料筆數共 16,003 筆。

再則，高雄 MaaS 系統會員電子票證搭乘紀錄同樣有資料龐雜且較不易切分之問題，雖從資料中可以清楚觀察到每一位會員的一筆交易紀錄，卻無法更明確得知在一天之內同一位會員於不同運具下的所有花費、次數等，故本研究亦將參考方案購買紀錄之篩選方式，同樣使用高雄 MaaS 系統會員註冊資料的識別碼欄位為串聯基礎，藉此彙整所有會員於每一日之電子票證搭乘紀錄；另考量不同運具間有著價格上的限制與差異，故會分別對捷運、市區公車、公路客運各自建立一套資料庫以利分析。

表 7　會員票卡交易紀錄資料欄位說明

持有方案	定　　義
0	未購買方案
1	公車暢遊方案 (一般)
2	公車+客運暢遊方案 (一般)
3	無限暢遊方案 (一般)
4	公車暢遊方案 (學生)
5	公車+客運暢遊方案 (學生)
6	無限暢遊方案 (學生)

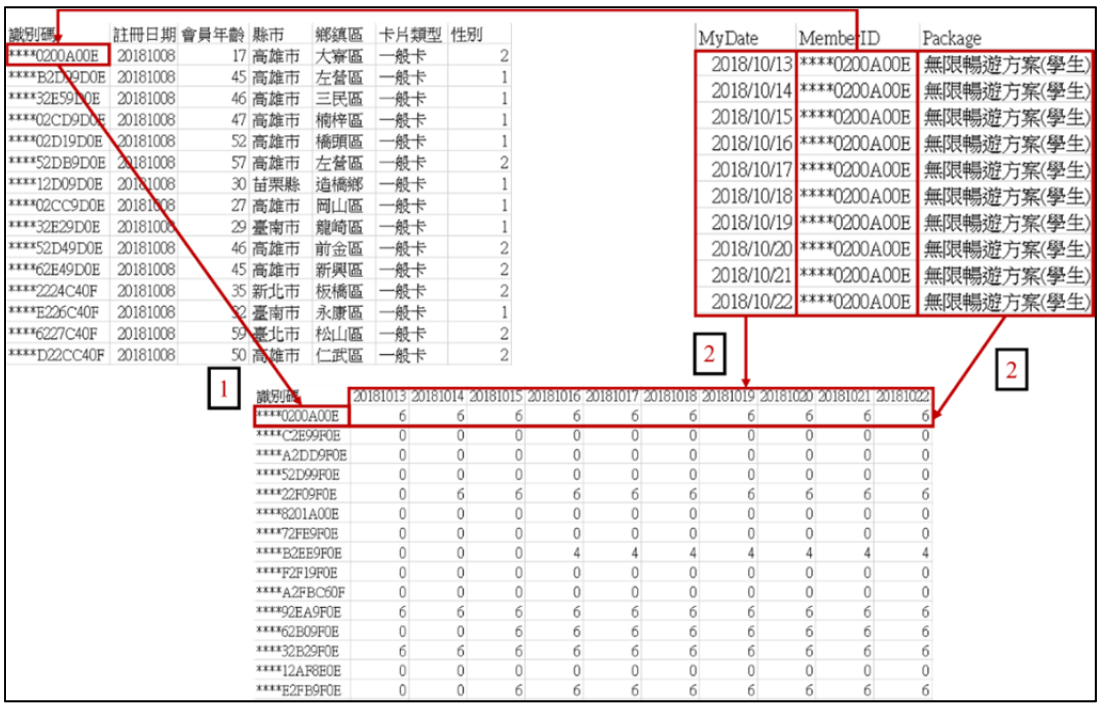


圖 7　方案購買紀錄資料勾稽示意圖

3. 欄位增生與變數轉換

經由資料勾稽後建立出一套逐月的會員方案購買紀錄資料庫，以及各運具於各月份的花費金額資料庫，本研究進一步透過計算上述資料庫欄位來推估每一位會員在每個月持有方案時的各運具總花費金額；另考慮到月份可能也是會員考量是否續買方案的因素之一，故增加一欄位來記錄購買方案當月之隔月月份。而除了新增欄位外，本研究也將會員註冊資料中的社經資料欄位進行轉換，轉換規則如表 8 到表 10 所示；其中，在一般卡及學生

表 8　一般卡年齡欄位轉換說明

年齡類別	定　　義
1	歲數 ＜18 (未成年)
2	歲數 ＝18 (剛成年)
3	19 ≦ 歲數 ≦ 22 (大學生)
4	23 ≦ 歲數 ≦ 45 (青壯年)
5	46 ≦ 歲數 ≦ 65 (中年)
6	65 ＜ 歲數　(老年)

卡之年齡欄位上，本研究將 18 歲自成一組，係考量 18 歲為使用者自僅能選擇大眾運輸轉換為可選擇私人運具之重要階段，故其方案購買行為可能與其他年齡層較為不同，如：不續買方案比例較大，若該特徵屬性 (18 歲) 得於分析時成為一分支條件，便代表本研究能進一步瞭解使用者於大眾運輸、私人運具轉換階段之方案購買行為，藉此推出適當之誘因以維持其繼續使用大眾運輸，對於未來推動 MaaS 亦能有所助益，而在居住地欄位上，為能簡化分類類別數量，故本研究將縣市、鄉鎮區欄位的部分則整合為「居住地」一欄位。

表 9　學生卡年齡欄位轉換說明

年齡類別	定　　義
1	歲數 ≦ 15 (國中小)
2	16 ≦ 歲數 ≦ 17 (高中)
3	歲數 = 18 (剛成年)
4	19 ≦ 歲數 ≦ 22 (大學生)
5	23 ≦ 歲數　(其他)

表 10　居住地欄位轉換說明

居住地類別	定　　義
1	高雄捷運有經過之行政區
2	高雄捷運未經過之行政區
3	臺南市
4	屏東縣
5	其他縣市

4. 處理後資料欄位定義

經前述步驟處理後之資料，各項資料欄位之變數類型與定義如表 11 所示。

5. 資料分群

由於 MaaS 會員資料具有一般卡與學生卡之身份別差異，也因此會造成方案購買選擇上的不同，為了避免後續建立之預測模型會有明顯不合理的分類錯誤，故需再將已處理後之資料依照會員身份將資料區分為一般卡資料以及學生卡資料等兩類別，並個別建立一套預測模型，以達到更合理、準確之預測效果。

6. 交叉驗證

本研究為確立決策樹的穩定性以及避免模型會有過度擬合的狀況，因此採用 5 折分層

交叉驗證 (stratified 5-fold cross validation) 以確認資料型態，其示意圖如圖 8 所示。交叉驗證的第一步驟為將資料樣本隨機分成 5 個資料集，第二步驟是將挑選 5 個資料集當中的 4 個資料集建立訓練資料集，且使用剩下的 1 個資料集作為測試資料集，第三步驟即重複第二步驟，然而不同的部分在於測試資料集的測試資料在 5 次的交叉驗證過程中皆只會當過一次測試資料集，就從此不再作為測試資料，直至所有資料集都曾被當作為測試資料集為止。而本研究使用分層交叉驗證是考量樣本資料具有資料不平衡的現象，若以學生卡方案購買狀況數量為例，從表 12 可以明顯看出來無限暢遊方案的數量為其他 3 種狀況的數倍，分層交叉驗證的目的就在於每一次的資料劃分中都能保持著原始數據中各個類別的比例關係。

表 11　處理後資料欄位定義說明

欄位名稱	變數類型	定　　義
Age	離散型	會員年齡
Live	離散型	會員居住地
Gender	離散型	會員性別
BusMonAvg	連續型	會員公車月花費
IntMonAvg	連續型	會員客運月花費
MRTMonAvg	連續型	會員捷運月花費
NextMonth	離散型	下個月之月份
下次所持方案	離散型	下個月所購買之方案 (目標變數)

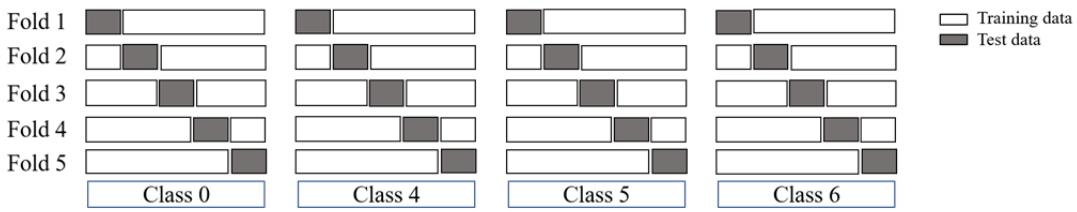


圖 8　5 折分層交叉驗證 (Stratified 5-fold cross validation)示意圖

7. 訓練與測試資料生成

在資料經過一般卡與學生卡的身份分群，且再透過 5 折分層交叉驗證分出訓練資料集與測試資料集後，為了可以再次驗證第三節分析結果中欲使用之 PDB 方法可優於 SMOTE

或只使用原始資料，因此後續分析之資料也將分別使用 PDB 方法、SMOTE 與原始資料，共 3 種型態的資料進行比較，以學生卡 Fold 1 為範例之樣本數量如表 13 所示。而為確保不同類別樣本間比較之公平性以及資料的真實性，本研究將使用 3 種不同類別樣本的訓練資料建立決策樹與支持向量機，並以相同的測試資料來觀察、比較各類別樣本之預測表現。

表 12　學生卡各方案購買情況數量

續買情況	數量
不續買 (0)	2430
公車暢遊方案 (4)	1300
公+客暢遊方案 (5)	425
無限暢遊方案 (6)	10470

表 13　學生卡 Fold 1 訓練資料各方案購買情況樣本數量範例

續買情況	樣本數量		
	原始資料	SMOTE	PDB
不續買 (0)	1942	8324	8324
公車暢遊方案 (4)	1050	8324	8324
公+客暢遊方案 (5)	339	8324	8324
無限暢遊方案 (6)	8324	8324	8324

五、實證分析

5.1 實驗環境及模式參數設定

本研究係利用機器學習套件 Scikit-learn 進行決策樹模型實測，並利用 Python 程式語言編寫程式碼，於搭載 Intel Core i5-8265U CPU、8GB RAM 之硬體環境下進行分析。針對本研究所使用之決策樹與支持向量機模型其設置及相關參數上，決策樹模型除 max_depth 以外，其他皆利用 Scikit-learn 之決策樹預設參數作為模型設置基礎 (Pedregosa *et al.* [23])，決策樹之各項參數設定如表 14。而本研究所用之支持向量機模型，則是利用 Scikit-learn 之 SVM 系列中的 SVC (C-Support Vector Classification)，並皆以 Pedregosa *et al.* [23] 建議之預設參數作為模型設置基礎。

表 14　決策樹相關參數設定

參　　數	說　　明
criterion	有 gini 或 entropy 兩類，entropy 因為有對數運算所以運算效率較低，一般而言都是使用預設的 gini 係數即可。
splitter	有 best 或 random 兩類，前者會在特徵所有劃分點中找出最佳劃分點，後者則是隨機在部分劃分點中找局部最佳劃分點，預設為 best。
max_depth	決策樹的最大深度，預設值為 None，即決策樹在建模時不會限制樹的深度。本研究經以數據測試後，將此參數設定為 6。
max_features	劃分時考慮的最大特徵數，有 log2 或 sqrt 兩類，預設值為 None，通常樣本特徵數若不多 (如小於 50)，就會選擇使用預設值 None。
min_samples_split	節點再劃分時所需最小樣本數，如果某節點的樣本數小於 min_samples_split 的設定值，則就不會繼續再繼續劃分，預設值為 2。
min_samples_leaf	每一節點的最小樣本數，如果某節點被劃分後的新節點樣本數小於 min_samples_leaf，則此新節點將不會出現，預設值為 2。

資料來源：Pedregosa *et al*. [23]；本研究整理。

5.2 績效指標說明

為了有效提升高雄市 MaaS 的營運績效，本研究建立會員方案續買預測模型，並以「月」為單位進行分析，期望能夠準確預測出高雄市 MaaS 有可能即將流失的會員，得以向目標會員精準投放促銷資訊及折扣。

圖 9 為本研究以一般卡 Fold 2 決策樹建立預測模型之範例，由分析結果可以得知類別 0 的 recall 為 0.77，代表預測模型對於下個月不會續買方案的會員 410 人中準確找出其中的 314 人，而剩餘的 96 人會被誤判為仍會續買方案的成員，故高雄市 MaaS 將會流失掉這 96 人；另一方面類別 0 的 precision 為 0.79，其中的 400 人是被預測模型預測下個月不會續買方案的會員，這 400 人裡真正下個月不會續買方案的會員有 314 人，而剩餘的 86 人則是被誤判為不會續買方案的成員，然而這 86 人其實下個月還是會續買方案，故高雄市 MaaS 會因為這個預測結果而將多支出額外的促銷、折扣成本在這些仍會續買方案的 86 人身上。所以根據研究與營運目的，類別 0 的 Recall、Precision 可以作為在 MaaS 會員方案續買預測上的績效指標，類別 0 的 Recall 和 Precision 值能夠越高越好，而 F1-score 是 Recall 和 Precision 的調和平均數，通常 F1-score 的值越高也代表 Recall、Precision 的值越高，故本研究最終將選擇使用在 classification report 上類別 0 的 Recall、Precision 以及 F1-score 成為在 MaaS 會員方案續買預測上的比較績效指標。

5.3 一般卡分析

表 15 為一般卡 3 類別樣本之訓練資料和測試資料數量分布情形，表 16 及表 17 則為

一般卡會員資料使用支持向量機與決策樹方法經過 5 折分層交叉驗證的平均預測結果績效指標。

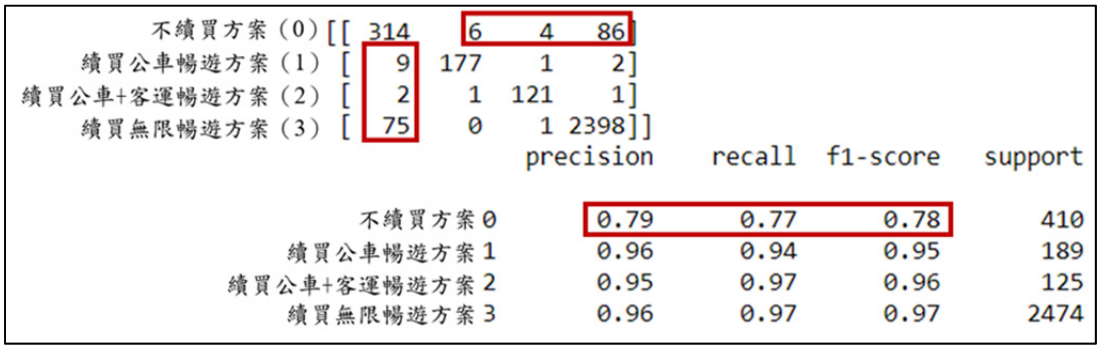


圖 9　一般卡 Fold 2 決策樹預測結果

表 15　一般卡每一次交叉驗證之資料數量分布

續買情況	數量			測試資料
	原始資料	SMOTE	PDB	
不續買 (0)	1633	9889	9889	409
一般卡 / 公車暢遊方案 (1)	760	9889	9889	191
一般卡 / 公+客暢遊方案 (2)	506	9889	9889	118
一般卡 / 無限暢遊方案 (3)	9889	9889	9889	2496

表 16　一般卡於 SVM 交叉驗證之測試資料績效指標平均結果

績效指標 / 訓練資料型態	Precision	Recall	F1-score
原始資料	**0.832**	0.676	0.744
SMOTE	0.680	**0.852**	0.754
PDB	0.812	0.824	**0.820**

根據表 16 之分析結果，在 5 次 SVM 交叉驗證的結果平均下來，PDB 方法的 F1-score 結果最好，而 Precision 和 Recall 的表現分別為原始資料與 SMOTE 最佳，但是其實 PDB 方法的值與其皆相差不到 3%，且原始資料的訓練資料雖能夠提升 Precision 的準確率，

SMOTE 的訓練資料能夠提升 Recall 的準確率，但也因此使其分別犧牲了 Recall 和 Precision 的表現，將導致增加誤判，使營運單位需要花不少不必要的支出在這些誤判上，也是這些不佳的預測表現上造成原始資料與 SMOTE 在 F1-score 上之表現和 PDB 方法相差至少 6%。

表 17　一般卡於決策樹交叉驗證之測試資料績效指標平均結果

績效指標 / 訓練資料型態	Precision	Recall	F1-score
原始資料	0.796	0.778	0.788
SMOTE	0.660	**0.852**	0.740
PDB	**0.842**	0.844	**0.844**

從表 17 之分析結果可發現，在 5 次決策樹交叉驗證的結果平均後，PDB 方法的 Precision 與 F1-score 結果最好，而 Recall 的表現為 SMOTE 最佳，但是其實 PDB 方法與 SMOTE 相差不到 1%，並且 SMOTE 的訓練資料雖能夠提升 Recall 的準確率預測出最多即將流失的會員，但也因此使其犧牲了 Precision 的表現，將導致許多仍會續買的會員被誤判為即將流失，使營運單位需要花不少不必要的支出在這些誤判上，也是 SMOTE 在 Precision 不佳的預測表現上造成其 F1-score 為 3 種型態的訓練資料下最差的結果。

而從交叉驗證的預測結果上可以看出，PDB 方法的訓練資料在 SVM 與決策樹下所建立的預測結果可以算是整體中表現最佳。本研究進一步以表格方式把高雄市 MaaS 一般卡會員決策樹的主要分支條件列出如表 18 所示。由彙整之分析結果可以發現，雖然一般卡會員只要客運月花費大於 280.177 元就會幾乎被歸類在公+客暢遊方案，但是絕大部分會續買公+客暢遊方案的會員主要客運月花費會大於 1440.341 元，而這也與實際高雄市 MaaS 系統所販售的公+客暢遊方案價格接近，不過這也代表會續買公+客暢遊方案的會員其花費行為與續買其他方案的會員花費行為有著明顯的差異，故才會發生續買公+客暢遊方案的會員主要客運月花費會大於 1440.341，但是只要會員的客運月花費大於 280.177 就會幾乎被預測續買公+客暢遊方案的情況，而這個現象也能在其他兩個方案上發現到。

高雄市 MaaS 一般卡會員決策樹除了上述的主要分支以外，也有並非花費相關的分支條件，其結果如表 19 所示。依照規則可以發現在相同的花費前提下，如果下個月為 1、2 月將會不續買方案，而其他月份會預測為續買無限暢遊方案，會有這個現象推測是因為這些會員在 12 月、1 月時購買方案已是 12 月、1 月中下旬，故其方案將能使用至 1 月、2 月多，而前者如果在 1 月後續買將會面臨 2 月份過年的情況，故才會在 1 月份就不續買方案，後者則是在 2 月多使用完方案時，除了會面臨到過年的可能，也還有月底的 228 連假，實際上 2 月份的上班日就不多，故買了時效內用的次數不多便也跟著不划算，也因此就不續買。

根據一般卡會員之決策樹分析結果可以發現，影響方案是否續買之主要影響因素為大眾運輸之月花費，當月花費到達一定程度且接近或超過套裝方案售價時，消費者便會選擇購買套裝方案，顯示套裝方案之方案內容及其售價將大幅影響會員之方案購買行為，若 MaaS 營運業者與政府單位為提升民眾使用大眾運輸之誘因，亦可參考分析結果作為調整方案售價之參考基礎。再則，月份為一般卡會員在大眾運輸花費以外之分支條件，而主要涉及不購買方案之月份多與春節連假有關，故 MaaS 營運業者或許可以在春節連假或其他連續假期月份，適當提供會員優惠以避免顧客流失之情形發生，抑或可思考於春節連假前後期間提供短期套票方案、輕量級使用方案進行販售，讓這些於 1、2 月份有需求但需求量不大的會員有適合的套票方案可以選擇，便不會因此放棄續買方案。

表 18　高雄市 MaaS 一般卡會員決策樹主要分支條件

續買情況	分支條件 (單位：元)
不續買 (0)	客運月花費 ≦ 280.177 捷運月花費 ≦ 947.928 公車月花費 ≦ 138.483
一般卡 / 公車暢遊方案 (1)	客運月花費 ≦ 280.177 捷運月花費 ≦ 947.928 公車月花費 > 138.483 (公車月花費主要 > 329.406)
一般卡 / 公+客暢遊方案 (2)	客運月花費 > 280.177 (客運月花費主要 > 1440.341)
一般卡 / 無限暢遊方案 (3)	客運月花費 ≦ 280.177 捷運月花費 > 947.928 (捷運月花費主要 > 1230.231)

表 19　高雄市 MaaS 一般卡會員決策樹特殊分支條件

續買結果	分支條件
不續買方案 (0) 或一般卡無限暢遊方案 (3)	客運月花費 ≦ 0.006 947.928 < 捷運月花費 ≦ 1230.231 NextMonth ≦ 2.5 (下個月是否為 1、2 月)

5.4 學生卡分析

表 20 為學生卡 3 類別樣本之訓練資料和測試資料數量分布情形，表 21 及表 22 則為學生卡會員資料使用支持向量機與決策樹方法經過 5 折分層交叉驗證的平均預測結果績效指標。

表 20　學生卡每一次交叉驗證之資料數量分布

續買情況	數量			測試資料
	原始資料	SMOTE	PDB	
不續買 (0)	1942	8324	8324	488
學生卡 / 公車暢遊方案 (4)	1050	8324	8324	262
學生卡 / 公+客暢遊方案 (5)	339	8324	8324	86
學生卡 / 無限暢遊方案 (6)	8324	8324	8324	2069

表 21　學生卡於 SVM 交叉驗證之測試資料績效指標平均結果

績效指標 / 訓練資料型態	Precision	Recall	F1-score
原始資料	0.796	0.744	0.768
SMOTE	0.672	**0.852**	0.75
PDB	**0.838**	0.772	**0.804**

表 22　學生卡於決策樹交叉驗證之測試資料績效指標平均結果

績效指標 / 訓練資料型態	Precision	Recall	F1-score
原始資料	0.762	0.81	0.784
SMOTE	0.684	**0.874**	0.768
PDB 方法	**0.798**	0.834	**0.814**

表 21 之分析結果中，在 5 次 SVM 交叉驗證的結果平均下來，PDB 方法的 Precision 與 F1-score 結果最好，而 Recall 的表現為 SMOTE 最佳，雖然 SMOTE 的訓練資料能夠提升 Recall 的準確率預測出最多即將流失的會員，但也因此使其犧牲了 Precision 的表現，將導致許多仍會續買的會員被誤判為即將流失，將導致增加誤判使營運單位需要花不少不必要的支出在這些誤判上，也是這些不佳的預測表現上造成原始資料與 SMOTE 在 F1-score 上之表現和 PDB 方法相差至少 5%。

在表 22 之分析結果方面，將 5 次決策數交叉驗證的結果平均後，PDB 方法的 Precision 與 F1-score 結果最好，而 Recall 的表現則是 SMOTE 最佳，PDB 方法與 SMOTE 相差 4%，

雖然 SMOTE 的訓練資料能夠提升 Recall 的準確率預測出最多即將流失的會員，但也因此使其犧牲了 Precision 的表現，將導致許多仍會續買的會員被誤判為即將流失，也是 SMOTE 在 Precision 不佳的預測表現上造成其 F1-score 為 3 種型態的訓練資料下最差的結果。

而從交叉驗證的預測結果上可以看出，PDB 方法的訓練資料在 SVM 與決策樹下所建立的預測結果可以算是整體中表現最佳。本研究同樣將高雄市 MaaS 學生卡會員決策樹的主要分支條件列出如表 23 所示，由彙整之分支條件可以發現，雖學生卡會員只要客運月花費大於 656 元就幾乎會被歸類在公+客暢遊方案，但是絕大部分會續買公+客暢遊方案的會員主要客運月花費會大於 1141.108 元，而這也與實際高雄市 MaaS 系統所販售的公+客暢遊方案價格接近，不過這也代表會續買公+客暢遊方案的會員其花費行為與續買其他方案的會員花費行為有著明顯的差異，故才會發生續買公+客暢遊方案的會員主要客運月花費會大於 1141.108 元，但是只要會員的客運月花費大於 1141.108 元就會幾乎被預測續買公+客暢遊方案的情況，而這個現象也能在其他兩個方案上發現。

表 23　高雄市 MaaS 學生卡會員決策樹主要分支條件

續買情況	分支條件(單位：元)
不續買 (0)	客運月花費 ≦ 656 捷運月花費 ≦ 761.611 公車月花費 ≦ 197.099
學生卡 / 公車暢遊方案 (4)	客運月花費 ≦ 656 捷運月花費 ≦ 761.611 公車月花費 > 197.099 (公車月花費大多 > 460.008)
學生卡 / 公+客暢遊方案 (5)	客運月花費 > 656 (客運月花費主要 > 1141.108)
學生卡 / 無限暢遊方案 (6)	客運月花費 ≦ 656 捷運月花費 > 761.611 (捷運月花費主要 > 1235.442)

高雄市 MaaS 學生卡會員決策樹除了上述的主要分支條件外，也有並非花費相關之分支條件，其結果如表 24 所示。首先，表 24 第一列依照規則可以發現在相同的花費前提下，當下個月份並非 1 月時 (NextMonth>1.5)，模型會預測為續買無限暢遊方案，而這時的標準為會員捷運月花費只要大於 761.611 元即可，然而當下個月份為 1 月時 (NextMonth≦1.5)，此時的會員需要捷運月花費需要大於 1590.093 元模型才會預測下個月續買無限暢遊方案，在此推測因為 2019 年 1 月多為學生開始放寒假的時刻，故在 12 月時預測模型將標準提高，12 月時對捷運通勤有足夠高度需求的會員，1 月時也才有機會在放寒假前有能夠通

勤消費達到回本的程度，甚至是這些會員有除了上學通勤以外的搭乘需求(例如去補習)，故即使放寒假的時候這些會員還是會有固定的通勤需求在。

再則，從第二列依照規則可以發現在相同的花費前提下，如果年齡小於 18 歲 (Age≦ 2.5) 的會員，下個月仍然會續買公車暢遊方案，而成年會員不會續買方案，然而此情況之月花費金額還不到學生卡公車暢遊方案的定價，故推測這一些未成年的會員因為還是高中生以下的學生身分，生活開銷由家中提供多，加上父母於平日也不方便親自接送小孩上下學，所以直接購買方案方便孩子自行通勤，故即使月花費不及方案價格也仍會持續購買方案以滿足孩子的通勤需求，所以也可以反推年齡大於 18 歲的會員生活所需的開銷可能需要自己全部負責或者部分負責，開始有經濟壓力在也會對金錢感到比較精打細算，因此當月花費不及方案價格時也理所當然不會讓自己賠本，就不會再續買方案而會去尋找其他更划算的運具使用。

針對學生卡會員之決策樹分析結果上，使用者不續買方案之主要分支條件同為大眾運輸之月花費，惟可發現除捷運月花費金額略低於一般卡會員外，客運、公車月花費之分支條件皆高於一般卡會員，顯示學生卡會員對於客運、公車之需求較高，未來亦可嘗試對學生族群提供更優惠之客運、公車暢遊方案以提升其購買意願。在大眾運輸月花費以外之分支條件主要有二，分別為月份及是否小於 18 歲，月份部分同樣與寒假、春節連假有關，故同樣可利用推出短期套票方案、輕量級使用方案作為行銷策略；另在是否小於 18 歲之分支條件上，由於本項分支條件尚可能與生活開銷是否需由使用者自身負擔有關，而 18 歲又大致可作為學生屬高中、大學之分界點，故未來亦可考慮根據學生之年級別推出不同之套裝方案，如：大學生因需自行擔負部分生活開銷，故可適當推出較優惠之價格或較實惠之大學生專屬方案內容以提升其購買意願，而高中生因無使用私人運具之選項，對於大眾運輸之需求較高，且加以考量高中生因多為家長負擔開銷，故高中生專屬方案價格便可較大學生族群略為提高，藉此將市場區隔做進一步劃分。

表 24　高雄市 MaaS 學生卡會員決策樹特殊分支條件

續買結果	分支條件
不續買方案 (0) 或學生卡無限暢遊方案 (6)	客運月花費 ≦ 656 捷運月花費 > 761.611 NextMonth ≦ 1.5(下個月是否為 1 月) 捷運月花費 ≦ 1590.093
不續買方案 (0) 或學生卡公車暢遊方案 (6)	客運月花費 ≦ 656 捷運月花費 ≦ 761.611 133.623 ≦ 公車月花費 ≦ 197.099 Age ≦ 2.5(年齡是否小於 18 歲)

六、結論與建議

本研究主要以資料探勘之決策樹方法建構一套高雄市 MaaS 系統方案續買預測模型，以 MaaS 系統提供之會員註冊資料、方案購買紀錄以及電子票證搭乘紀錄進行資料處理，並將資料使用 PDB 方法、SMOTE 與原始資料來產生三種類型之資料作為輸入資料，進行決策樹方法模型之訓練及探討各類型資料下所建立之決策樹的預測結果，最終經由 5 折分層交叉驗證後發現 PDB 方法產製之資料所建出之決策樹預測模型在一般卡別與學生卡別都有不錯的預測成效，以下茲說明本研究之主要成果並提出未來相關研究之建議。

6.1 結論

1. 本研究透過勾稽高雄市 MaaS 系統所記錄之會員註冊資料、方案購買紀錄以及電子票證搭乘紀錄，可依研究分析課題需要產製出資料探勘方法所需之訓練資料，且能具體應用於預測 MaaS 會員之方案購買行為。
2. 為能改善 SMOTE 方法可能導致資料型態改變，以及加劇資料點分散情況之缺點，本研究以資料之原始分佈型態為切入點，提出以機率分配為基礎之增加抽樣方法來增加少數類別樣本，透過以網路公開資料集進行測試與比較，亦發現本研究提出之 over-sampling 方法明顯優於 SMOTE 方法，顯示 PDB 方法除具有學術面之貢獻外，亦具有實務上應用之可行性。
3. 本研究根據 PDB、SMOTE 等兩種 over-sampling 方法產生分析資料集，並建立 MaaS 方案購買行為之預測模型，根據 5 折分層交叉驗證後發現，以 PDB 方法生成訓練資料所建置出的決策樹與支持向量機模型，無論在一般卡、學生卡皆有較佳的預測結果。在一般卡會員方案續買預測模型中，PDB 方法之績效指標分數達 81.2%到 84.4%，學生卡會員方案續買預測模型之績效指標分數則達 77.2%到 83.8%。
4. 透過觀察決策樹的分支條件可發現，影響方案是否續買之主要影響因素為大眾運輸之月花費，故 MaaS 營運業者與政府單位亦可視政策需求，以調整定價方式提升民眾使用大眾運輸之誘因。
5. 除大眾運輸月花費之分支條件外，一般卡與學生卡皆出現月份成為分支條件之情形，其中又以寒假、春節連假之 2 月份出現不續買方案之行為最明顯，故建議 MaaS 營運業者可適當於春節連假前後期間提供短期套票方案、輕量級使用方案，藉此令使用者可持續購買方案並提升大眾運輸使用率。
6. 根據決策樹分析結果，本研究亦發現學生卡會員在 18 歲以上、18 歲以下之方案購買行為不盡相同，當中又以 18 歲以上學生對於方案購買之性價比較為在意，故本研究亦推測此情況與大學生相較於高中生，可能有更多機會需自行擔負生活開銷有關，故建議 MaaS 營運業者未來可進一步將高中以下學生與大學生之方案內容進行區分，藉此令套票方案可更符合不同學生族群之使用需求。

6.2 建議

1. 本研究所使用之會員註冊資料、方案購買紀錄以及電子票證搭乘紀錄僅使用 2018 年 11 月至 2019 年 5 月共 7 個月份的資料量，若未來有機會能夠取得更多時間的資料，應能建立一套更加完整之方案續買預測模型。
2. 本研究所使用的資料因為電子票證搭乘紀錄之部分資料缺失，如：下車時間、上下車站名，而使得資料欄位的增生程度有限，若能有更完整的資料，將可以有助於更深入瞭解 MaaS 系統使用者的套票購買行為。
3. 本研究所使用之變數仍以運具花費為主，雖然已加入部份社經變數，但決策樹模型主要分支變數仍是以運具花費為大宗，因此若能試著投入更多非運具花費的變數或許可以有更多樣化的分支條件出現，並能提出更豐富之行銷策略建議供相關單位參考。
4. 高雄市 MaaS 系統於 2020 年、2021 年間新推出許多月票優惠價購買方案，本研究根據一般卡會員與學生卡會員決策樹主要分支條件所提出之改善建議，亦能與上述所提到之優惠方案相呼應，故後續研究亦可進一步觀察新推出之月票優惠方案是否有助於提升會員之方案續買意願，也能藉此驗證本研究分析之方案購買預測結果與實際方案購買行為是否相符。
5. 本研究利用不續買方案之主要、特殊分支條件提出若干方案調整、新增建議，惟研究結果之解釋多係築基於先驗知識而得，雖討論內容對於一般使用者而言尚屬合理，惟業者後續仍能透過問卷調查蒐集使用者不續買之考量因素，藉此驗證本研究在分析結果之驗證、解讀上是否正確，並可作為後續研究深入探討方案購買行為之參考。

參考文獻

1. 盧宗成、王晉元、簡佑勳、楊煜民、王蕾絜、林季萱、吳東凌、陳翔捷，「交通行動服務會員之共享運具使用特性分析–以高雄市 MeN Go 系統為例」，**運輸計劃季刊**，第 51 卷，第 3 期，民國 111 年，頁 195-230。
2. 廖振宇，「應用資料探勘技術於公車間轉乘策略之研究」，淡江大學運輸管理學系運輸科學碩士論文，民國 104 年。
3. 鍾智林、李舒媛，「以悠遊卡大數據初探 YouBike 租賃及轉乘捷運行為」，**都市交通**，第 33 卷，第 1 期，民國 107 年，頁 16-36。
4. 劉芷璇，「MATLAB 應用於公車乘載率分析-以臺中市 35 路公車為例」，逢甲大學運輸科技與管理學系碩士論文，民國 105 年。
5. 林至康、張志鴻、蘇昭銘、張朝能、沈美慧、蔡欽同，「運用電子票證資料推估大眾運輸旅次訖點之演算法構建與驗證」，**運輸計劃季刊**，第 47 卷，第 1 期，民國 107 年，頁 1-28。
6. Long, Y. and Thill, J. C., “Combining Smart Card Data and Household Travel Survey to

Analyze Jobs–Housing Relationships in Beijing" , *Computers, Environment and Urban Systems*, Vol. 53, 2015, pp. 19-35.

7. Morency, C., Trepanier, M., and Agard, B., "Analysing the Variability of Transit Users Behaviour with Smart Card Data", 2006 IEEE Intelligent Transportation Systems Conference, 2006, pp. 44-49.
8. 林浩瑋，「悠遊卡大數據應用於大眾運輸乘客旅運型態之研究」，淡江大學運輸管理學系運輸科學碩士論文，民國 105 年。
9. Rodríguez, J., Semanjski, I., Gautama, S., and van de Weghe, N., "Unsupervised Hierarchical Clustering Approach for Tourism Market Segmentation Based on Crowdsourced Mobile Phone Data" , *Sensors*, Vol. 18, No. 9:2972, 2018.
10. Meng, C., Cui, Y., He, Q., Su, L., and Gao, J., "Travel Purpose Inference with GPS Trajectories, Pois, and Geo-Tagged Social Media Data," 2017 IEEE International Conference on Big Data, 2017, pp. 1319-1324.
11. 王晉元、盧宗成、李晟豪、陳其華、吳東凌、陳翔捷，「手機信令資料探勘於改善觀光旅客公共運輸服務之研究－以花蓮縣臺灣好行路線為例」，**運輸計劃季刊**，第 48 卷，第 2 期，民國 108 年，頁 105-131。
12. 賴盈臻、邱裕鈞，「基於行動信令資料之使用運具判斷模式」，**運輸學刊**，第 33 卷，第 3 期，民國 110 年，頁 285-309。
13. Ramboll, Whimpact, *Insights from the World's First Mobility-as-a-service (MaaS) System*, MaaS Global, Denmark, 2019.
14. Wong, Y. Z., Hensher, D. A., and Mulley, C., "Emerging Transport Technologies and the Modal Efficiency Framework: A Case for Mobility as a Service" , 15th International Conference Series on Competition and Ownership in Land Passenger Transport, 2018, Collected Papers.
15. Barua, S., Islam, M. M., Yao, X., and Murase, K., "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning" , *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 2, 2012, pp. 405-425.
16. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-Sampling Technique" , *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 2825-2830.
17. 陳威廷、張永佳，「應用增生少數合成技術與二階段邏輯斯迴歸方法建構汽車租賃信用風險評估模型」，交通大學運輸與物流管理學系碩士論文，民國 106 年。
18. 王銘亨，「閃光號誌路口交通事故特性分析-以臺南市為例」，**交通學報**，第 16 卷，第 1 期，民國 105 年，頁 39-54。
19. Chen, M. C., Chen, L. S., Hsu, C. C, and Zeng, W. R., "An Information Granulation based Data Mining Approach for Classifying Imbalanced Data" , *Information Sciences*, Vol. 178, Vol.16, 2008, pp. 3214-3227.
20. 陳世承，「不平衡資料集學習之少數類別過抽樣技術的一個改良方法」，清華大學資訊工程學系所碩士論文，民國 106 年。
21. Malerba, D., UCI Machine Learning Repository - Page Blocks Classification Data Set,

website: https://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification, Retrieved January 27, 2020.

22. UCI, UCI Machine Learning Repository, , website: http://archive.ics.uci.edu/ml, Retrieved January 27, 2020.
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., “Scikit-learn: Machine Learning in Python” , *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825-2830.