題目：消費者族群購物周期及商品預測

李瑞庭、徐悅寧、常思為、劉馨蔓

摘要

本研究主要目的在於為 PChome 創造更大的商業利益，帶來更多的顧客消費，而在顧客從進入了 PChome 平台的瞬間，便要透過不同面向來分析顧客的一舉一動，首先若能知道顧客進入 PChome 的時間點，歸納出顧客出現在平台的週期，讓我們能夠預期不同時間點的差異所要應對的顧客數量會是千軍或是萬馬，以備後台支援的軟硬體、人力及物力。

而當顧客開始進行選購商品的動作時，我們便也開始找出商品間的關係是否環環相扣，選了第一件商品便再找出具關聯的商品推薦給顧客，讓顧客購物車裡如同聚寶盆一般不計其數。

顧客確認是否遺漏要買的東西時，我們可利用他的消費記錄預測出當次他購買的商品及金額。最終，當顧客點下結帳鍵，踏進了收銀區後，便產生了一筆交易紀錄豐富了我們的數據庫，這筆資料能夠帶給我們未來進行預測，利用分析來看他是否為我們的忠誠顧客，並且預測是否為潛在 Prime 會員的身份。

關鍵詞：關聯分析、機器學習、精準行銷、主成份分析、RFM

壹、前言

一、研究背景

2020 年初，全球經歷一個前所未有的風暴——新冠肺炎，在疫情的肆虐之下人們的生活也隨之產生了莫大的變化。對於 PChome 而言，在這場疫情帶來最大的挑戰便是顧客的消費趨勢改變，衝擊了原本以「24 小時到貨」、「台北市 6 小時到貨」贏得大眾青睞的 PChome，因為訂單量暴增，尤其生活用品、食用品訂單如沙塵暴般襲捲而來，造成後台無法負荷短時間內破百萬的下單次數、商品供應量不足，以至於發生出貨延遲的問題。這樣嚴重的問題使 PChome 大失去了長年累積的商譽，也讓喜愛 PChome 的顧客對 PChome 的忠誠度有了變化。

二、資料來源

2021 年經濟部工業局舉辦 data station 數據競賽 PChome 組，由 PChome 提供。資料為 PChome 19998 位顧客在 2020 年的消費訂單，共 1027855 筆，變數有消費者編號、郵遞區號 、訂單編號、訂單日期、訂單時間、商品類別、商品名稱、商品編號、訂單商價和 Prime 卡會員。

貳、分析方法

在本研究中，我們做了四種分析方法，分別是測消費者購買商品週期、商品回購預測、Prime 潛在會預測、關聯分析，希望能透過統計分析、機器學習等方法為 PChome 增加商業利益，在做分析前，我們先對四種分析方法中前處理皆會使用的 RFM 模型

(一) RFM 模型介紹

利用 PChome 企業提供的交易紀錄,將顧客的資料以同日期同時間視為一筆交易，刻劃出 Recency(R)代表顧客最後一筆購買紀錄距離基準點的天數，Frequency(F)代表顧客的交易筆數，Monetary(M)代表顧客的總消費金額。

(二) 預測消費者購買商品週期

為了瞭解 PChome 的消費者購買商品週期，使 PChome 業者能在特定的時間宣傳商品，以達到精準行銷，降低宣傳成本，達到最高的利益，我們將使用機器學習方法預測每位消費者是否在指定的時段內進行消費。此報告以預測消費者是否會在 30 天購買為例。

1. 依照執行步驟

(1) 消費者所在地區切割資料：

由於在台灣不同地區的消費者可能會有不同的消費習慣，我們依照 PChome 資料中的郵遞區號變數來切割資料，並分別切割成北部地區資料集、中部地區資料集、南部地區資料集。北部地區包含地區有台北市、基隆市、新北市、宜蘭縣、連江縣、新竹市、新竹縣、桃園市、苗栗縣；中部地區有台中市、彰化縣、南投縣、嘉義市、嘉義縣、雲林縣；南部地區包含台南市、高雄市、澎湖縣、金門縣、屏東縣、台東縣、花蓮縣。為了評估本次報告所使用的機器學習模型之效能，尤其是已訓練完成的模型在新資料上的表現，減少過度擬合的可能性，我們將北部、中部、南部、臺灣全地區依照時間切割資料，並切割成 3 份資料集，表 1 說明訓練資料集分別為 1 到 3 月、5 到 7 月、9 到 11 月交易紀錄資料；測試資料集分別為 4 月、8 月、12 月消費者是否購買。

表 1 切割資料方法

	自變數使用資料	依變數
第一份切割資料	1 到 3 月交易紀錄資料	4 月消費者是否購買
第二份切割資料	5 到 7 月交易紀錄資料	8 月消費者是否購買
第三份切割資料	9 到 11 月交易紀錄資料	12 月消費者是否購買

(2) 使用的變數：

I. 自變數：

我們將第一、二、三份切割資料分別探索其各自的 RFM 資訊，並將各自的 R、F、M 資訊分別使用正規化處理、正規化處理後使用 Kmeans 分群的分群數，因此共有六個自變數，並分別獲得第一、二、三 RFM 資訊資料。

II. 依變數：

本分析方法是一分類問題，分別將第一、二、三份切割資料中測試資料集分成兩類，在指定月份中，消費者有購買商品行為及消費者沒有購買商品行為。

(3) 模型配適：

我們將訓練及測試資料比例取 8：2，使用五種機器學習模型，分別為 Random Forest Classifier(RF) 、 Multilayer Perceptron Classifier(MLP) 、 eXtreme Gradient Boosting Classifier(XGBoost)、Light Gradient Boosting Machine Classifier(LightGBM) 、Logistic Regression Classifier(LR)進行配適。

(三) 商品回購預測

每個人都有屬於自己的消費習慣，因此我們想了解消費者該次消費是否會對商品進行回購，以刺激消費、提升利潤，我們將整理資料型態，並使用機器學習方法，預測消費者此次上線購物時，是否會對曾購買過的商品進行回購。

以下皆以衛生紙為例。

執行步驟

(1)整理資料：

由於原始資料是以所有交易商品作為主體，我們希望將資料型態整理成一個 dataframe，當中只提取曾經買過衛生紙的消費者的所有訂單交易資訊，並由原始資料新增變數，變數有以下四種：

I. 該商品上次購買至現在的時間長

II. 商品 30 天內的購買頻率

III. 顧客購買商品時為星期幾

IV. 購買的時段(一天以每 4 小時劃分)

表 2 購買過衛生紙的人的訂單為主體的 dataframe

編號	時間間格	購買頻率	星期幾	購買時段
0	12	0	1	3
1	94	0	4	3
2	39	0	1	6
3	10	0	4	5
4	6	0	3	3

(2)模型配適：

由於當中被有購買衛生紙的資料和未購買衛生紙的資料，比例相差巨大，因此我們採用半監督式學習(Semi-supervised Learning)的方式，搭配下採樣(undersampling)隨機取出 10000 筆衛生紙交易資訊與 15000 筆非衛生紙交易資訊進行機器學習模型的訓練，由於單一模型訓練不佳，因此我們使用 KNN、Logistic Regression、Random Forest、MLP、LGBM、XGBoost 六個模型預測，並從中挑選適合該資料的模型，只要其中 2 個模型判斷該特徵為會購買衛生紙我們及判定為會購買衛生紙，並重複抽樣 5 次取得平均。

此想法是參考 Cumby, Fano et al. 2004 所提出的論文，認為曾經買過衛生紙的人有高機率會重複購買，且他們也許會有相似的消費習慣，可由機器學習模型提煉。

(四) Prime 潛在會員預測

在 PChome 這次題目的敘述中，我們得知會員的制度對於 PChome 的重要性，這些具備 Prime 卡片的會員所具備的特徵為何？如何鞏固這些忠誠會員才能加強對顧客特質的掌握？

方法一執行步驟：

(1)使用的變數：

I. 自變數

A. 消費者購買間隔、頻率、消費金額以 Recency, Frequency, Monetary 共三項表示之。

B. 將消費者郵遞區號資料進行分類：分別為 北部、中北部、西部、西南部、南部、東部共六項。

C. 計算消費者訂單中個類別數量，包含 3C、日用、生活、休閒、衣鞋包錶、周邊、美妝、食品、家電、書店、通訊、筆電、數位共十三項。

II. 依變數：

以 0 表示非會員，1 表示會員

(2)預測方法：

Naive Bayes Classifier（Naive Bayes ）、K-Nearest Neighbors Algorithm（KNN）、Random Forest Classifier（RF）、Logistic Regression（LR）、Decision Tree Classifier（DT）、eXtreme Gradient Boosting（XGBoost）、Adaptive Boosting（AdaBoost）、Gradient Boosting Classifier（Gradient Boost）、Histogram-based algorithm（Hist）、Stacking Classifier（Stacking）

方法二執行步驟：

(1)使用的自變數：

I. RFM：Recency, Frequency, Monetary

II. 地區：北部、中北部、西部、西南部、南部、東部

III. 商品類別：3C、日用、生活、休閒、衣鞋包錶、周邊、美妝、食品、家電、書店、通訊、筆電、數位

(2)降低維度：

使用主成份分析將資料降低到三維度。

(3)預測方法：

將降低維度後的資料以散佈圖呈現在空間上，並根據已知的是否為會員區分出來，若非會員的資料散佈在會員中密集的區域者，判定其為潛在會員。

(五) 關聯性分析

縱使每個人的消費習性不甚相同，但是經過分析仍會有相似之處，就如同著名的尿布啤酒理論。在這次 PChome 所提供的資料中，我們想要從中發掘某些特定的族群是否會購買相同類型的商品組合。經由分析結果，我們能對會一同購買的商品提供優惠組合，藉此來刺激消費，提升營收。

- 支持度(Support)：在所有訂單中，同時購買商品 A 和商品 B 的訂單占所有訂單的比例。
- 置信度(Confidence)：在所有購買商品 A 訂單中，同時購買商品 A 和商品 B 的比例。
- 提升度(Lift)：置信度/支持度。

  提升度< 1，表示商品 A 的出現和商品 B 的出現是負相關。

  提升度> 1，表示商品 A 的出現和商品 B 的出現是正相關。

  提升度= 1，表示商品 A 和商品 B 為完全獨立個體無相關。

方法一執行步驟：

(1) 商品分類：

取商品 id 前四碼(例如：商品 DAAL07-A9008O0WF-000 取前四碼 DAAL)作為商品分類依據，並從中選取較常出現的商品 id 前四碼（例如：衛生紙）作為欲分析關聯性的商品，將有購買衛生紙的購物清單取出並作分析。

(2) 分析：

使用 python 套件 apyori 中的 apriori 指令對衛生紙做關聯性分析。

方法二執行步驟：

(1)商品分類：

透過 tfidf 找出常出現的字詞，並從中選取是商品的字詞來做分析，將含有該字詞的商品的購物清單取出，作為欲分析的對象。

(2)分析：

透過條件機率計算，當消費者該筆消費購買衛生紙時，同時購買其他商品的比例。

分析結果

(一) 預測消費者購買商品週期

由於北、中、南地區消費者購買商品數量懸殊，由圖 1 可發現，已北部消費者購買商品數量最多，高達 800000 多項商品，而中、南部消費者購買商品數量則落於 100000 項左右，由此可知北部消費者的購買力相較於中南部是非常驚人的。

北部與中南部地區消費者不僅購買力差異大，消費頻率也不盡相同，由表 4 可知，北部與中南地區消費者在 30 天內重複購買商品的比例差異也大，中南部消費者在 30 天內重複購買商品的比例是北部地區消費者的 5 倍。

由上述可知，各地區消費者的購買習慣各有不同，為了增進預測消費者購買週期的準確率，我們將消費訂單依照消費者所在地區劃分成 3 份資料，分別進行模型配適，也為了比較將消費者依所在地劃分資料與不依照地區劃分資料的配適效果是否有差異，我們也使用全台灣地區消費者交易訂單進行模型配適。

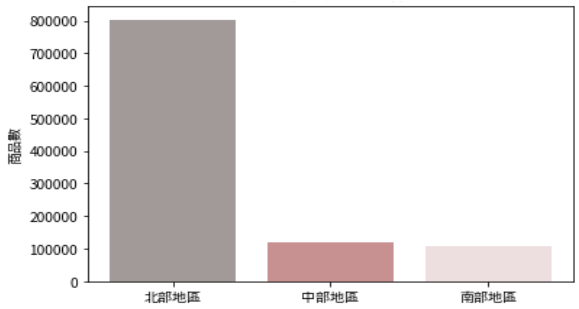

圖 1 各地區消費者購買商品數量長條圖

表 3 在 30 天內重複購買商品的消費者占當地區的比例

北部地區	中部地區	南部地區	全部地區
0.0002	0.0010	0.0011	0.0001

我們分別使用機器學習中的 Random Forest、Multilayer Perceptron Classifier、eXtreme Gradient Boosting Classifier、Light Gradient Boosting Machine Classifier Logistic Regression 五種模型進行配適，表 5 為以北、中、南部地區、台灣全地區在三個指定期間(1-3 月、5-7 月、9-11 月)訂單資料所獲得的 RFM 正規化、RFM 正規化後使用 KMeans 分群數結果配適五種模型所得的準確率及召回率的平均，圖 2 則是以表 5 為資訊的雷達圖，由(a)可知，依照地區畫分消費者訂單後，配適模型所得平均召回率結果明顯比不依照地區劃分消費者訂單高，然而由(b)可發現平均準確率結果則適依照地區畫分消費者訂單後比不依照地區劃分消費者訂單低，兩者的結果迥異。

表 4 使用機器學習模型配適資料所得結果

地區	模型	準確率	召回率
北部地區	LR	0.665	0.807
	MLP	0.654	0.798
	SVM	0.615	0.876
	XGBoost	0.614	0.732
	CATBoost	0.665	0.809
中部地區	LR	0.670	0.755
	MLP	0.596	0.552
	SVM	0.714	0.832
	XGBoost	0.616	0.678
	CATBoost	0.709	0.839
南部地區	LR	0.849	0.701
	MLP	0.689	0.670
	SVM	1.000	0.674
	XGBoost	0.810	0.663
	CATBoost	0.825	0.701
台灣全地區	LR	0.850	0.720
	MLP	0.819	0.717
	SVM	1.000	0.631
	XGBoost	0.804	0.681
	CATBoost	0.863	0.711

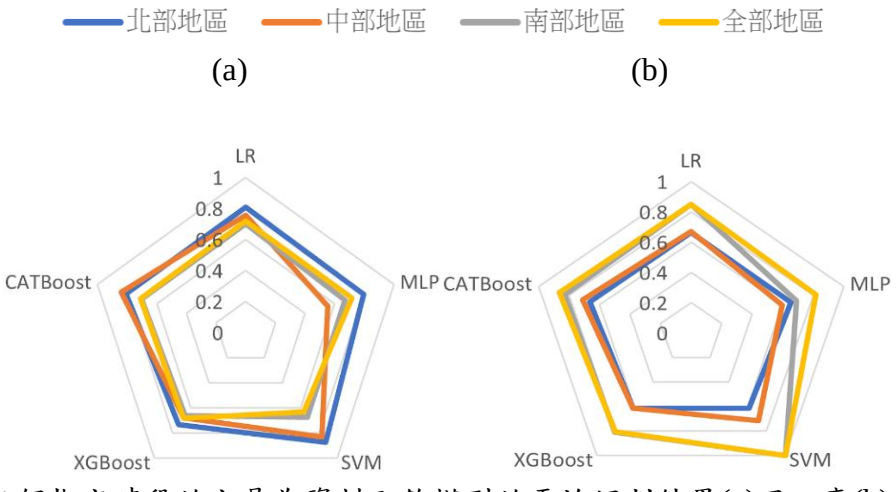

圖 2 三個指定時段的交易為資料配飾模型的平均評判結果(a)召回率(b)準確率

在此分析方法中，召回率相較於準確率是重要的，因為在此的召回率定義為預測出 30 天會購買的顧客中，有多少占比的顧客會真正在 30 天內再次消費，PChome 業者可以針對預測會再次購買商品的人進行宣傳，促進他們再次消費，若針對少部分並非會購買的顧客宣傳，也不會造成太大的損失。

圖 2(a)中可知，若是依照地區將顧客交易資訊劃分並各自配是模型，可以獲得較高的平均召回率。圖 2(a)中也可知，北部、中部、南部、台灣全地區獲得的最高平均召回率模型分別為 SVM、CATBoost、CATBoost、Logistic Rregression。

下公式為使用台灣全地區顧客交易資訊並使用邏輯式回歸(Logistic Rregression)的計算方法，由公式可知顧客的交易筆數(F)正規化變數、總消費金額(M)正規化變數式較為重要的變數。

$$ln\frac{p}{1-p} = -0.36 - 1.00X_1 + 144.37X_2 + 13.73X_3 - 0.26X_4 - 0.05X_5 - 0.73X_6$$

其中 p 為顧客在 30 天內購買的機率；X_1為顧客最後一筆購買紀錄距離基準點的天數(R)正規化變數；X_2為顧客的交易筆數(F)正規化變數；X_3為顧客的總消費金額(M)正規化變數；X_4為顧客最後一筆購買紀錄距離基準點的天數(R)正規化並用 KMeans 分群分群數變數；X_5為顧客的交易筆數(F)正規化並用 KMeans 分群分群數變數；X_6為顧客的總消費金額(M)正規化並用 KMeans 分群分群數變數

(二) 商品回購預測

表 5 機器學習單一模型結果

模型	準確率	召回率
KNN	0.885	0
LR	0.885	0
RF	0.7348	0.3533
MLP	0.115	1
LightGBM	0.8495	0.2404
XGBoost	0.8711	0.1879

由上表可觀察可知，KNN 和 LR 模型訓練的結果會把所有的觀察值判斷為「此次消費不會購買衛生紙」，MLP 訓練的結果則會認為所有觀察值為「此次消費會購買衛生紙」，因此我們推測 KNN、LR、MLP 等模型不適合配適此資料，而 Tree-based 的模型 RF、LightGBM、XGBoost 在此資料有較好的表現。

機器學習合併模型結果：

我們組合 Random Forest、LightGBM、XGBoost 此三模型，只要其中一個模型判斷資料為「此次消費會購買衛生紙」我們則判定他會進行購買。

表 6 機器學習合併模型結果

模型	準確率	召回率
RF+LGBM+XGB	0.59	0.6

(三) Prime 潛在會員預測

使用方法一，以機器學習方式預測，結果如下：

表 7 Prime 潛在會員以機器學習預測結果

模型	訓練資料召回率	測試資料召回率
Naive Bayes	0.1978	0.5953
KNN	0.6616	0.6298
RF	0.6992	0.6715
LR	1.0000	0.4701
DT	0.6016	0.5898
XGBoost	0.6929	0.6715
Adaptive Boost	0.6544	0.6606
Gradient Boost	0.6795	0.6661
Hist	0.6596	0.6497
Stacking	0.7678	0.6733

下方公式為使用顧客交易資訊並使用邏輯式回歸(Logistic Rregression)進行預測潛在會員的計算方法：

$$\ln\left(\frac{p}{1-p}\right) = -0.00264 - 0.0062X_1 + 0.0041X_2 + 0.0000061X_3 - 0.0060X_4 - 0.05X_5 - 0.73X_6 + 0.0016X_7 + 0.0051X_8 - 0.0038X_9 + 0.0020X_{10} - 0.0087X_{11} - 0.00035X_{12} - 0.0053X_{13} + 0.0053X_{14} + 0.017X_{15} - 0.0016X_{16} - 0.000031X_{17} - 0.00072X_{18} - 0.0011X_{19} + 0.0024X_{20} + 0.0031X_{21} + 0.017X_{22}$$

其中 p 為顧客是潛在會員的機率；X_1為顧客最後一筆購買紀錄距離基準點的天數變數；X_2為顧客的交易筆數變數；X_3為顧客的總消費金額變數；X_4為顧客購買 3C 類產品數量變數；X_5為顧客購買休閒類產品數量變數；X_6為顧客購買周邊類產品數量變數；X_7為顧客購買家電類產品數量變數；X_8為顧客購買數位類產品數量變數；X_9為顧客購買日用類產品數量變數；X_{10}為顧客購買書店類產品數量變數；X_{11}為顧客購買生活類產品數量變數；X_{12}為顧客購買筆電類產品數量變數；X_{13}為顧客購買美妝類產品

數量變數；X_{14}為顧客購買衣鞋包錶類產品數量變數；X_{15}為顧客購買通訊類產品數量變數；X_{16}為顧客購買食品類產品數量變數；X_{17}為顧客於北區購買次數變數；X_{18}為顧客於中北區購買次數變數；X_{19}為顧客於中區購買次數變數；X_{20}為顧客於中南區購買次數變數；X_{21}為顧客於南區購買次數變數；X_{22}為顧客於東區購買次數變數。

由此回歸公式可知每一個係數皆很小，計算出來的會員機率會很低，預測效果是裡面方法最不佳的，因此我們傾向於使用 Random Forest Classifier、XGBoost Classifier 及 Stacking Classifier 等方法搭配使用已獲得較佳的預測能力。

使用方法二，以空間中的分佈圖預測，結果如下：

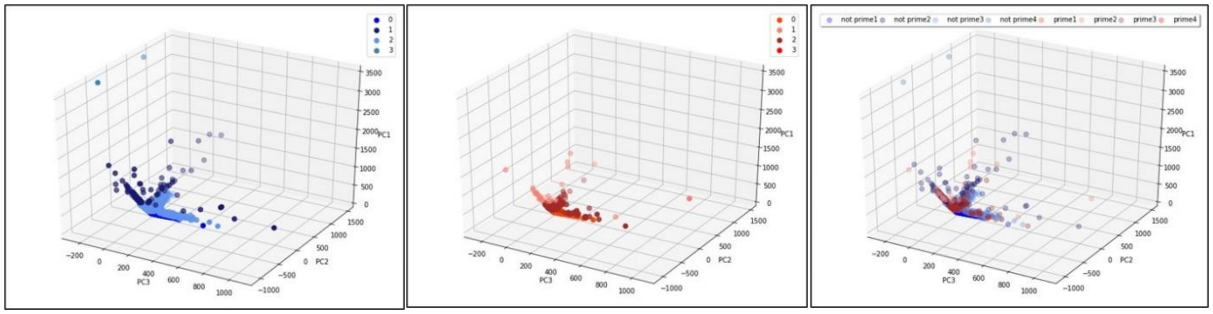

圖 3 以空間中的分佈圖預測三維圖（左）使用非 Prime 會員訂單資料（中）使用 Prime 會員訂單資料（右）使用全部訂單資料

在最密集的區域也就是第一主成分介於（-80~ 80），第二主成分介於（-40~0），第三主成分介於(-40~0)當中，我們回測 2020 年的資料，預測潛在會員能力達到 69%。

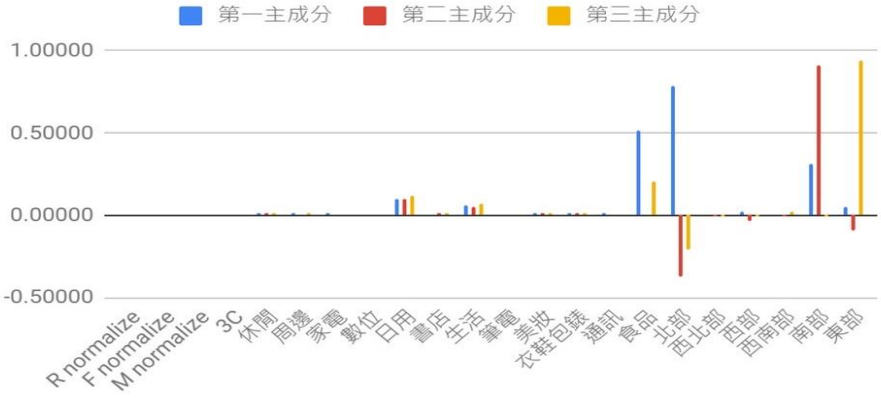

圖 4 變數降維後各係數權重

(四) 關聯性分析

方法一：

我們使用在 Fan, C.-K. (2003)的論文中有提及的支持度、信賴度，我們使用在 Python 中的套件 apriori，這是專門用於做關聯性分析，因此我們透過 apriori 對衛生紙做關聯性分析的結果如下表。可發現消費者在購買衛生紙的同時，也會有高機率

購買礦泉水、罐頭、進口食材和速食麵。但在置信度方面皆趨近於 1，意思是當消費者購買衛生紙時，有高達 98%的機率會購買礦泉水、罐頭、進口食材和速食麵，但是經過驗證查詢後，發現消費者購買衛生紙時，同時購買礦泉水、罐頭、進口食材和速食麵的機率並沒有那麼高，因此透過 apriori 對衛生紙做關聯性分析的結果是錯誤的，我們推斷利用 apriori 做關聯性分析時應該將所有的消費紀錄放進套件中分析，找出商品的關聯性，而不是由特定的商品清單去對該商品找關聯性，但是 PChome 所提供的資料比數過大，無法透過 apriori 做關聯性分析，因此我們不採用此種方法所得出的結論。

表 8 衛生紙 apriori 關聯性分析結果

商品	支持度	置信度	提升度
礦泉水	0.0227	0.9801	1.0122
罐頭	0.0242	0.9832	1.0154
進口食材	0.0216	0.9833	1.0155
速食麵	0.0221	0.9836	1.0159

方法二：

透過條件機率計算得出購買衛生紙時同時購買的商品前四名分別為洗衣精、衛生棉、家用清潔劑和零食，比例約莫都落在 3%上下。

表 9 衛生紙條件機率結果

商品	比例
洗衣精	0.0363
衛生棉	0.0310
家用清潔劑	0.0309
零食	0.0306

結論

我們針對上述分析方法中的預測消費者購買商品週期、商品回購預測、Prime 潛在會預測、關聯分析做出結論。

(一) 預測消費者購買商品週期

由分析結果可知,由於地區消費者消費型態不盡相同,若是依照地區劃分消費者,獲得各地區消費者購買力 RFM 特徵,並依照 RFM 特徵進行機器學習模型配適,預測消費者是否於 30 天內重複購買商品,可以發現預測結果中的平均召回率相較於直接使用全臺灣地區消費者 RFM 特徵為資料提升許多,由此可知,「地區」特徵對於消費者消費型態有相當大的影響力。

(二) 商品回購預測

此部分是要預測商品是否回購,能找出有衛生紙的訂單是主要目標,因此衡量指召回率的重要程度大於準確率,召回率值 0.6 意味著,會買衛生紙的人我們能準確的預測出 6 成,對於電商的效益十分巨大,也許該名消費者有購買衛生紙的意願,但他準備結帳時衛生紙不在購物籃中,這時我們推薦衛生紙給他,就有很高的機率能增加獲利,在 Cumby, Fano et al. 2004 論文當中也指出,用此方式幫助芝加哥的連鎖量販店增加了 11%的獲利,因此我們將我們的商品回購預測與 Cumby, Fano et al. 2004 論文中的結果進行比較。

表 10 Cumby, Fano et al. 2004 提出的結果與我們的結果

	準確率	召回率
random	0.65	0.19
sameas	0.70	0.26
top-10	0.59	0.37
Perceptron	0.65	0.38
Winnow	0.79	0.17
C4.5	0.77	0.22
Hybrid-Per	0.53	0.59
Hybrid-Win	0.65	0.43
Hybrid-C4.5	0.62	0.46
Our method	0.59	0.6

文獻中作者最後要呈現的方法為最後三列 Hybrid 的結果,與我們的成果比較分數相當接近,準確率互有優劣,但我們的方法召回率比論文高,因此我們認為結果是成功的,原因我們猜測因該論文是 2004 年出版的,而我們使用的機器學習模型比當年厲害許多。我們也有做尿布、奶茶等產品,結果和衛生紙接近,準確率和召回率都在 0.5~0.7 之間。

(三) Prime 潛在會員預測

在進行主成份分析降低維度的步驟之後，我們發現消費者所在的地區跟食品、日用品對於消費者消費習慣的差異，有很大的影響力。

將資料投影到三維的空間中，我們發現會員的點會集中在某個區域，我們的想法是從 Prime 會員密度最高的地方尋找，尋找裡面的非 Prime 會員，在這個區域裡的非會員顧客，我們便將其視為潛在會員，透過這個想法我們對 2020 年進行回測，可以找出到 69%的潛在 Prime 會員。

以兩種方法作為評估是否為潛在會員，召回率的表現都能高達接近七成，在方法二的預測能力又更高了點，可見在會員當中我們能準確的預估出約七成的人。

並且從上述的分析中我們發現了重要的結果：顧客是否習慣購買食品以及顧客是否住在北部地區，對於潛在的 Prime 會員有很大的影響力。

在我們對於忠誠顧客的需求量大的情況下，我們希望能夠集結更多屬於潛在會員的顧客，因此，在兩種分析方法中，皆可以視預測出來為潛在會員的顧客為我們這次希望得到的潛在顧客，得到此資訊之後便可以再後續對這些顧客進行精準行銷，也可以對於顧客有更深層的洞察。

(四) 關聯性分析

以衛生紙為例，透過條件機率，當顧客購買衛生紙時也會去購買洗衣精、衛生棉、家用清潔劑和零食相較於其他商品高，比例約莫都落在 3%上下，因此當顧客選擇了衛生紙時，即可推薦上述四種商品給他，並給予優惠，刺激消費，替 PChome 創造更多的效益。

參考文獻

1. Predicting customer shopping lists from point-of-sale purchase data. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. (Cumby, C., et al.，2004).
2. 關聯法則與統計分析之探討, National Central University. (Fan, C.-K. ，2003).

附錄

表 11 各商品館所銷售之單件商品金額

商品館	總金額	平均數	標準差	最小值	中位數	最大值
全部	904694678	880.18	3041.55	1	293	165980
家電	109279308	3379.49	5980.18	12	1299	120799
休閒	45604509	932.32	2184.50	1	500	97500
日用	102352932	402.08	601.90	1	199	30000
筆電	45984394	22033.73	18358.84	99	21900	158940
通訊	129794768	2933.88	7287.48	9	545	122900
周邊	50593337	1024.01	2566.88	1	4755	143996
食品	99142966	316.90	466.17	1	179	23790
生活	101148790	736.23	2100.58	1	340	100000
美妝	22591429	634.59	822.24	1	396	34902
書店	6303873	372.97	757.55	16	277	23380
衣鞋包錶	22916838	743.72	1626.00	11	446.5	89999
數位	37683902	3265.50	7039.17	51	1470	163213
3C	80006493	3241.36	6126.78	6	1399	165980