Journal of the Chinese
Statistical Association
Vol. 63, (2025) 234–275

H&M 數據驅動的全通路客戶分析：RFM 於購買行為預測研究

張維欣　蘇南誠[†]

國立臺北大學統計學系

摘　要

現今企業必須善用客戶資料進行分析與預測，從而更好地了解客戶特徵，才能最大化客戶經營效益，並提供良好的全通路服務體驗。本研究利用 H&M 於 Kaggle 平台上發布的競賽資料，以交易資料計算 RFM 指標，採用集群分析 K-means 方法將其進一步分群，並且依全通路、線上或線下通路各別計算，同步探討使用不同長度間的交易記錄對模型預測能力的影響。藉由統計分析和機器學習等方法，預測客戶未來 1 個月和 3 個月的購買行為，並分類客戶為一般客戶和高貢獻客戶。根據本研究結果，建議企業善用 RFM 分群識別客戶，關注近期購買且高貢獻的客戶，並運用模型優化行銷策略，以 LightGBM 和邏輯斯迴歸預測高機率購買或潛力高貢獻的客戶。最後，除了數據應用，企業也應重視與客戶的長期互動，並優化數位體驗。

關鍵詞：決策樹、K-means、LightGBM、邏輯斯迴歸、全通路、RFM、XGBoost。

JEL classification: C45, M31.

[†]通訊作者：蘇南誠
E-mail: sunanchen@gmail.com

1. 緒論

2020 年起 Covid-19 疫情蔓延全球，各國紛紛實施不同程度的防疫措施，這些措施對消費者行為產生了顯著影響。在此背景之下，電子商務成為人們生活中不可或缺的一部分。根據 Chevalier (2022) 發表於 statista 的市場報告顯示，2020 年零售業的電子商務銷售額相較於 2019 年大幅增長了 26.8%。該報告還預測，即使在 2022 年後疫情逐漸趨緩，零售業的電子商務成長率仍保持在 9.7% 左右，詳細數據見圖 1。這表明消費習慣已經產生重大轉變，消費者可能會於實體門市嘗試產品或服務，轉身於網路上搜尋商品資訊及使用心得，最後因價格或便利性而選擇線上購物。因此，企業必須透過顧客關係管理，深入了解消費者需求，在對的時間點提供最適合的產品和服務。這仰賴透過客戶資料進行分析與預測，從而更好地了解客戶特徵，才能最大化客戶經營效益，提供良好的全通路（Omni-channel）服務體驗。

疫情不僅改變了消費者行為，同時也成為全球企業數位轉型的催化劑。Kotler *et al.* (2021) 提出了運用行銷科技的五種策略，策略之一為「資料行銷」，其主要目標為建立一個穩定的資料生態系，透過蒐集多樣化的數據來源，包含顧客社群資料、網路流量資料、交易資料、物聯網和互動資料等等。另一重要策略為「預測行銷」，運用機

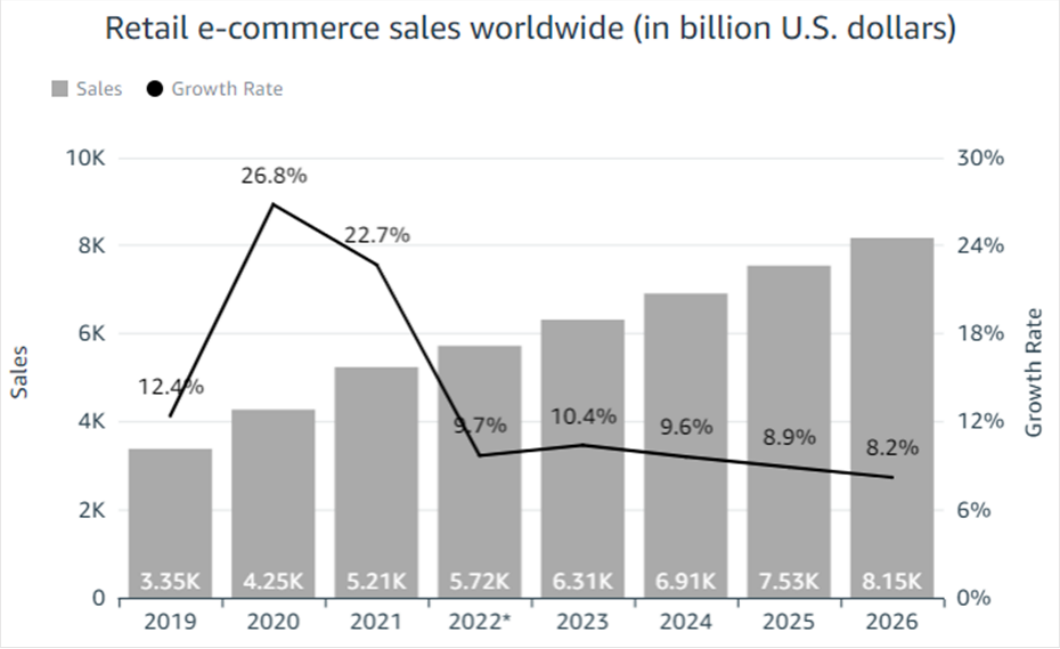

圖 1：全球零售業之電商銷售額與成長率。
資料來源：Chevalier (2022)。

器學習等分析技術，我們能夠從龐雜的數據中找出特定模式，行銷人員能夠憑此技術，預測哪些客群可能會消費或哪類活動可能會奏效等。

然而，建立完整的資料生態系統並非每個企業都能輕易實現，尤其面臨第三方 cookie 消失的議題，追蹤顧客的數位軌跡日漸困難。對於尚處於轉型或布局中的企業，更需要一套簡單可行且有效的分析方法，用於顧客經營以創造價值，為企業帶來實質的支援，促進其數位轉型過程的順利進行。

2022 年全球性集團 H&M 號召各路好手，協助以數據推敲客戶需求與喜好。H&M 是 Hennes & Mauritz 的縮寫，於 1947 年成立於瑞典，如今為全球性的時尚與設計公司，擁有超過 4,000 家店鋪，遍布 70 多個市場，並在 60 個市場開展線上銷售。H&M 主打平價快速時尚，產品組合包括女裝、男裝、童裝、鞋類、配飾、化妝品和家居用品。Smith (2023) 的市場報告中指出 H&M 在全球服裝零售商排名中排名第六，正由於 H&M 在全球擁有如此廣泛的市場和多樣化的業務範疇，對數據分析的需求尤為迫切與重要。數據分析不僅能夠協助 H&M 更精準地瞭解各個市場的消費者需求和行為模式，從而優化產品組合和存貨管理，提升銷售效益和顧客滿意度；更能夠深化顧客關係管理，透過對顧客購買行為和偏好的洞察，發展精準行銷策略和推廣活動。

H&M 集團包含多個品牌和業務，其線上商店為消費者提供了豐富多樣的商品選擇。然而，選擇過多可能會導致顧客無法迅速找到他們感興趣的或正在尋找的商品，最終可能不會進行購買。為了提升購物體驗，商品推薦顯得至關重要，協助顧客作出正確選擇甚至能減少退貨，從而減少運輸所產生的碳排放。因此，2022 年 H&M 於 Kaggle 平台上發布競賽「H&M Personalized Fashion Recommendations」(https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations)，邀請參賽者基於顧客、商品以及交易等數據開發商品推薦。

本研究利用此份競賽資料，包括商品基本資料、客戶基本資料和全通路的交易記錄等數據，透過統計分析和機器學習等方法達成以下目的：

1. 預測客戶未來購買行為：預測未來 1 個月和 3 個月的是否購買。

2. 區分客戶類型：基於預測結果，分類客戶為一般客戶和高貢獻客戶。

我們探討使用不同長度間的交易記錄，是否對於模型預測能力有所影響，以期找出最佳組合。此外，以交易資料計算 RFM 指標，分別為最近一次消費（Recency）、消費頻率（Frequency）、消費金頭（Monetary）的縮寫，採用集群分析 K-means 方法將其進一步分群，以區分客戶及作為顧客特徵，並且依全通路、線上或線下通路各別

計算。我們期望為這筆資料提供有效的分析及預測方法，發掘出消費者的購買模式、喜好和行為，協助企業更深入了解顧客行為和需求，並支援短期促銷活動和季度行銷計劃的執行，從而優化顧客關係管理，提升企業競爭力和客戶滿意度，此模式亦能延伸至其他零售相關產業。

本研究使用 Python 進行資料預處理、分析及模型建置與預測，搭配使用 AWS QuickSight 進行資料視覺化。接下來第二章探討顧客經營與 RFM 應用等相關文獻；第三章介紹集群分析與分類預測的機器學習方法；第四章為實證分析，先說明資料來源與變數處理過程，經敘述性統計分析，再以機器學習方法建立預測模型；最後，第五章針對研究成果歸納結論與建議。

2. 文獻回顧

2.1 顧客經營管理

隨著全球化和數位化的步伐加速，市場環境正經歷著前所未有的變革。在這樣情境下，消費者的需求和期望也日益多元化，使得企業在維護顧客關係、提供個人化服務以及拓展市場時面臨更多挑戰。如今，理解並滿足個別客戶的需求已成為企業塑造競爭優勢的關鍵，而過往產品導向的策略也必須轉換到以客戶需求為中心的策略。因此，顧客關係管理（Customer Relationship Management；CRM）逐漸從一個選擇轉變為企業成功的必要條件。

Kumar and Reinartz (2018) 在他們 Customer Relationship Management: Concept, Strategy, and Tools 一書中提到，顧客關係管理的目標是通過最有效地向客戶提供價值和滿意度，並從這種交流中提取商業價值，以獲得長期的競爭優勢。因此，對客戶及其偏好的了解對整個組織來說是極為重要的。從這個立場出發，顧客關係管理是一個策略性過程，它涉及選擇企業能夠最有利可圖地服務的客戶，並塑造公司與這些客戶之間的互動。最終目標是優化客戶目前和未來對公司的價值。此外，該書亦提到客戶價值的概念對於 CRM 至關重要。它指的是客戶關係對公司的經濟價值，表達為貢獻邊際或淨利潤。作為一個行銷指標，客戶價值不僅能評估行銷的有效性，也是一個重要的決策輔助工具。公司可以通過將客戶價值作為其決策過程的核心來測量和優化其行銷努力。考慮到客戶價值的概念，我們可以將 CRM 描述為分析和使用行銷數據庫，並利用通信技術來確定能最大化每個客戶對公司終身價值的企業實踐和方法。客戶價值能細分致許多種類，例如過往客戶價值（Past customer value；PCV）、客戶終

身價值（Customer Lifetime Value；CLV）、客戶推薦價值（Customer Referral Value；CRV）、客戶知識價值（Customer Knowledge Value；CKV）等等，不同的價值指標有其獨特應用場景及意義。

本國亦有眾多客戶價值相關研究，如鄭子萱（2021）以流行服飾品牌的會員與交易資料，進行顧客流失與顧客終身價值之預測。他們透過調整計算預測目標的時間長度來進行長短期模型表現的測試，並比較行銷隨機模型、簡單機器學習模型與業界常用的直覺模型的預測能力；其研究結果顯示在探討顧客流失時，長短期的行銷隨機模型與簡易機器學習模型均優於直覺模型，其中以隨機森林模型表現最佳。在顧客終身價值預測方面，短期各模型並無顯著差異，但行銷隨機模型在長期預測上更具靈活性，更能反應出顧客終身價值。最後依據預測結果，篩選出前 20% 的高價值顧客，使其更具行銷應用價值。

前述研究使用相同的歷史資料，預測未來不同時間點的客戶行為，然而當今企業資料呈現爆炸式增長，我們意識到隨著時間的推移，早期數據的預測價值可能遞減，新進的數據往往更能捕捉消費者行為的變化。因此，探討歷史數據的選擇策略為本研究著墨的重點之一，期望能夠在保持計算效率的同時，最大化模型對未來行為的預測能力。

2.2 RFM 模型

RFM 模型由 Hughes (1994) 提出，是一種用於分析顧客價值和顧客分群的技術工具，可衡量客戶忠誠度與價值，因其概念簡單、容易理解和實作，至今仍於企業界中廣為使用。RFM 模型包含三項指標：

1. 最近一次消費（Recency）：指顧客自最近一次購買以來的時間。一般來說，最近購買的客戶更有可能在未來再次購買，反之時間越久則顧客再次購買機率越低。

2. 消費次數（Frequency）：指顧客在分析期間內的消費次數。一般來說，購買次數越高代表對業者的滿意度與忠誠度越高，且更有可能在未來再次購買。

3. 消費金額（Monetary）：指顧客在分析期間內的消費總金額。一般來說，金額越高代表顧客的貢獻度越高，有助於企業識別高價值的客戶。

結合這三項維度，RFM 模型可以幫助企業對其顧客進行分群，從而制定更具針對性的市場策略。Hughes (1994) 設計的分群方式，是以每一個指標排序後均分五等份，依其價值最高者給予 5 分、最低者為 1 分，合併三項指標來看共有 125 群，舉例來

說，最有價值的客群為 555，而客群 111 則為最低價值的一群。此方法簡單直觀，不過可能因均等切分而忽略指標的分佈特性，如顧客集中於特定區間中則無法被突顯。

近年 RFM 相關研究，經常搭配集群分析方式，甚至依資料特殊屬性加以調整模型。比如程美蘭（2018）的研究，考量基金產業特性，以 RFM 模型為基礎，調整為 RFMC 擴充模型（R-Retention 留存率、F-Frequency 交易頻率、M-Monetary 平均申購金額、C-Contribution 顧客貢獻度）。運用 K-Means 對顧客進行集群分析，建立忠誠度及貢獻度兩構面的顧客價值行銷策略矩陣。接者以卡方檢定識別了不同客戶群的特徵，使企業深化顧客理解，並在資源有限的情況下，針對類型客戶制訂客製化行銷策略，提升顧客貢獻度和企業收益。

而蕭維嘉（2020）聚焦消費者在不同銷售通路的購買偏好及模式，運用零售服裝公司訂單，根據通路分群並計算 RFM 指標等變數。研究結果發現，消費者在實體通路的每次購買平均金額較高，可能由於門市店員的引導及前往門市的時間與精力考量，促使一次性購足的行為；跨通路消費者展現較高的品牌忠誠度，呈現更佳的 RFM 指標表現；相對地，雙通路消費者對價格較敏感，品牌應深化其與消費者的情感連結。總體而言，虛擬通路購買傾向正逐年增加，反映出消費者購物模式的轉變。

結合程美蘭（2018）與蕭維嘉（2020）的研究，消費者的購物通路偏好反映了其不同的購買動機；運用 RFM 指標進行客戶分群有助於企業根據客群特性制定差異化的經營策略。因此，本研究借鑑並融合這兩種方法，不僅使用 RFM 指標進行集群分析，還根據通路對部分關鍵變數進行細分，以期發展更精確的行銷策略。

3. 研究方法

本研究使用 H&M 的競賽資料，分析流程整理如圖 2。首先，我們界定了特徵與標籤的計算時間範圍，並依區間執行特徵工程與預測標籤計算。特徵工程包括計算 RFM 指標及運用 K-means 方法進行分群；而標籤計算則是產出兩項預測標籤：是否購買和貢獻分類。隨後，本研究採用了四種機器學習演算法建構預測模型，並比較模型之間的成果表現。詳細內容在本章接續的三小節中展開，第一節探討 K-means 分群方法及分群數量的選擇策略，第二節深入介紹所選用的各種演算法，最後一節則會闡述模型評估的標準。

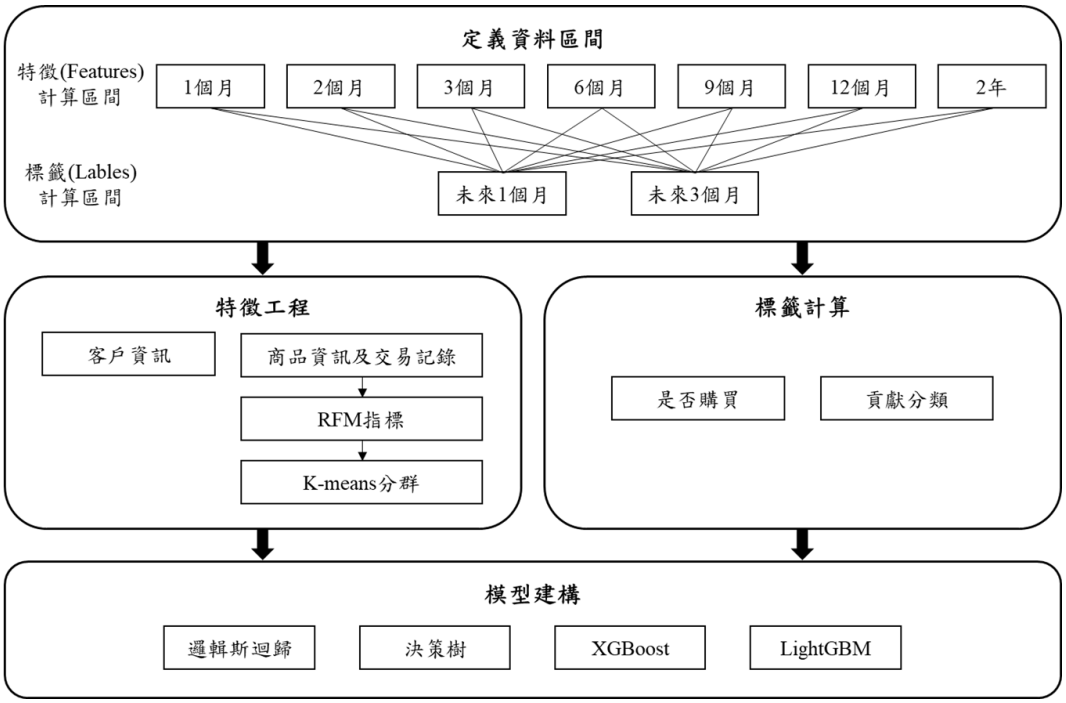

圖 2：分析流程。

3.1 K-means

為了以 RFM 指標進行客戶分群，我們採用集群分析方法，此方法用於將資料點按照其特徵的相似性分為多個群組，以確保相同群組內的成員具有高度相似性，而不同群組間之間則相對疏離。為了評估分群的準確性，Rousseeuw (1987) 提出了輪廓係數（Silhouette Coefficient），其值介於 -1 到 1，用來衡量群組的內聚力和分離度。當輪廓係數越接近 1 時，說明分群效果越理想，群組內的資料點非常緊密，並且與其他群組有著明顯的距離；相反，若輪廓係數接近 -1，則可能意味著某些數據點被誤分到了該群組，因為它們與其他群組的中心點更為接近。

K-Means 是一種廣泛使用的集群分析方法，其核心目標在於將資料分為指定的 K 個集群，並使得每個資料點被分配到最鄰近的集群中心（即平均值所在位置）。此算法以其直觀性和在處理大型數據集時的高效率而著稱，儘管其對初始的中心點敏感，可能導致最終解為局部而非全局最優，不過像是郭瀚揚（2019）和夏梅雪（2021）的研究結果，使用 K-Means 仍然能夠有效識別出不同特性的顧客群體。Anitha and Patil (2022) 的研究同樣使用了 K-Means 算法，並且以重複試驗的方式找出平均輪廓係數最高的 K 個集群。

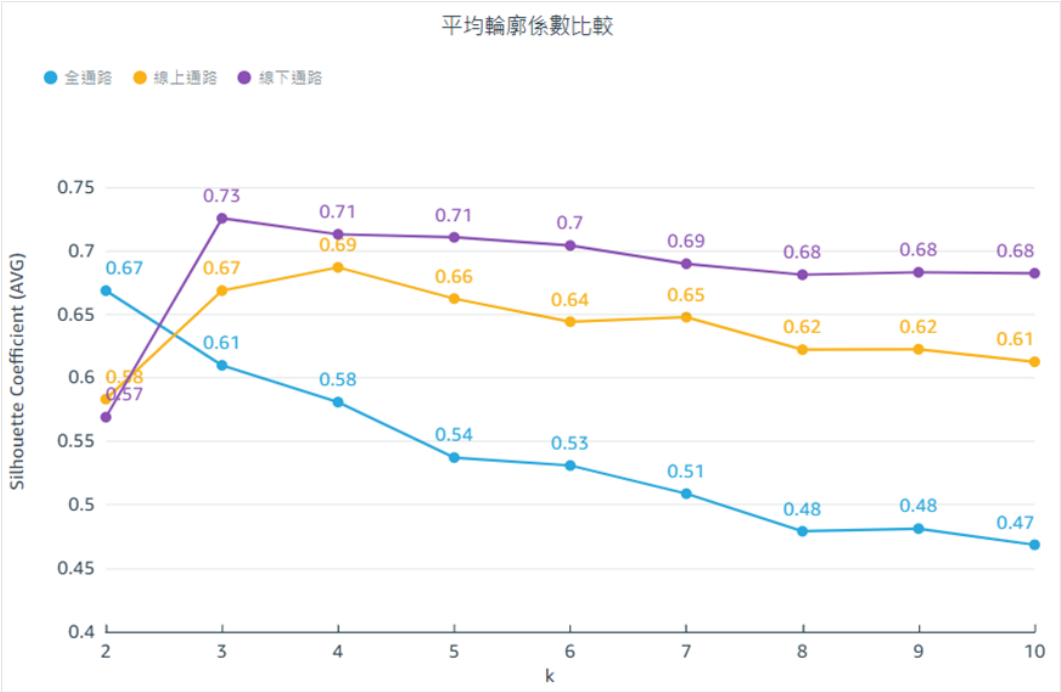

圖 3：K-means 分群之平均輪廓分數。

本研究選用了 K-Means 算法進行分群，主要因為資料集規模龐大，並且為了選擇最佳 K 群，以涵蓋完整年度的 12 個月特徵資料集隨機抽樣 10000 筆進行試驗，群數設定從 2 到 10，同步比較依全通路、線上和線下通路 RFM 指標分群後的平均輪廓係數，結果如圖 3。我們觀察全通路、線上以及線下通路的最高平均輪廓係數，分別出現在 $K = 2$、$K = 4$ 及 $K = 3$ 的分群設定。鑑於線上通路交易筆數佔比高達 70%，且是本研究重點探討的通路，我們決定以線上通路在最高平均輪廓係數下的 $K = 4$ 作為正式分群的基準。

3.2 分類模型

本研究的預測目標為客戶是否購買，並進一步分類客戶為一般或高貢獻客戶。為此，我們精選了適合分類任務的機器學習演算法，本節依序介紹：邏輯斯迴歸、決策樹、XGBoost，以及 LightGBM。每種模型都有獨到之處和適用情境，以下探討它們如何運作。

3.2.1 邏輯斯迴歸

邏輯斯迴歸（Logistic Regression）主要作為數據分析和推論工具使用，其目的是理解輸入變量在解釋結果中的作用（Hastie *et al.* 2009）。與傳統的線性迴歸模型不同，邏輯斯迴歸不是直接預測輸出值，而是透過 Sigmoid 函數預測特定結果的機率，這是一個能將任何實數投影在 0 到 1 之間的函數。

假設應變數 Y 為二元資料，對於給定的一組自變數 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$，令 $P(Y = 1|\boldsymbol{X}) = p$，邏輯斯迴歸模型公式如下：

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\,,$$

其中 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ 表係數，此模型的參數估計通常使用最大概似估計法（Maximum Likelihood Estimation；MLE）。一般來說，如果預測機率大於 0.5，則結果類別為 1，反之則為 0，因此結果類別為 1 的機率

$$P(Y = 1 \mid \boldsymbol{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}\,。$$

由於邏輯斯迴歸的模型簡單、易於實現以及解釋性強，廣泛應用於各種領域，從醫學診斷到市場行銷，再到金融風險評估等。

3.2.2 決策樹

決策樹（Decision Tree）亦是分類問題中的經典解決方案之一，它以樹狀結構來表示數據中變量間的決策規則和可能的結果，以其直觀和易於理解的特性而受到推崇。在決策樹中，每個內部節點代表一個特徵或屬性，每個分支代表這個特徵的一個可能值，而每個葉節點則對應於一個預測結果或類別。樹的建立從一個根節點開始，透過不斷地選擇最佳特徵來分割數據，這個過程通常是遞迴進行的。選擇特徵和分割點的標準包括資訊增益、基尼不純度和減少方差等，這取決於決策樹的具體類型，如 ID3、C4.5 或 CART 等。在本研究中，我們選用了基於 Breiman *et al.* (1984) 提出的 CART 算法進行優化的 Python 套件。

決策樹的主要優點是模型的解釋性強，可以直觀地展示數據是如何被分類的。此外，決策樹不需要對數據進行太多預處理，例如不需要標準化。然而，它們也有缺點，例如容易過擬合，特別是當樹變得特別深或複雜時。為了解決這個問題，通常會採用剪枝技術，如預剪枝和後剪枝，來限制樹的成長或簡化樹結構。

總的來說，決策樹因其簡單性和高解釋性，在各種領域，如金融信用評估、醫療診斷、市場分析等領域得到了廣泛應用。它們也是許多複雜機器學習算法，如隨機森林和梯度提升機的基礎。

3.2.3 XGBoost

延續前述對於決策樹的討論，梯度提升（Gradient Boosting）是一種進階的機器學習技術，由 Chen and Guestrin (2016) 提出，它基於決策樹的思想，但通過結合多個弱決策樹模型來構建一個更強大、更精確的預測模型。在這種框架中，XGBoost（eXtreme Gradient Boosting）是近年流行和強大的梯度提升變體之一。

XGBoost 的核心思想是逐步地建立決策樹，每棵新樹都試圖糾正前一輪模型的預測錯誤。這是通過應用梯度下降算法來最小化預測和實際值之間的差異實現的，這也是「梯度提升」名稱的來源。XGBoost 對傳統梯度提升方法進行了多項優化，如引入了正則化項來控制模型的複雜度，從而降低過擬合的風險。此外，XGBoost 還具有高效的計算性能和靈活性，支持多種目標函數和評估準則，使其成為處理大規模數據集的理想選擇。它提供了豐富的可調參數，如樹的深度、學習速率和剪枝策略等，允許使用者根據具體問題進行細緻調整。

由於其出色的性能和靈活性，XGBoost 在各種機器學習競賽和實際應用中獲得了廣泛的使用，特別是在複雜的分類和迴歸問題中表現優異。從金融信用評分到生物信息學的基因分類，XGBoost 都展現了其強大的數據建模能力。

3.2.4 LightGBM

最後，我們要討論的是 LightGBM（Light Gradient Boosting Machine），這是另一種高效的梯度提升框架，由 Ke *et al.* (2017) 提出，它在設計上針對更高效的計算性能和較低的記憶體消耗進行了優化。與 XGBoost 相比，LightGBM 在處理大規模數據時展現出更快的訓練速度和更低的記憶體使用。

LightGBM 和 XGBoost 同樣採用了基於直方圖的決策樹算法。這種方法將連續特徵值映射到一系列的離散箱中，然後在樹的構建過程中利用這些箱值而非原始值。這種做法大幅減少了計算分割點的次數，從而提升了計算效率。不同的是，LightGBM 還實施了葉優先的樹成長策略，在大數據集上可以更快地減少損失。除了高效能的計算表現外，LightGBM 還提供了許多便利的功能，如原生支援類別特徵，意即無需事先轉換類別變數，以及自動處理遺漏值等，這使得在實務應用中部署 LightGBM 變得

更加方便且強大。

由於其卓越的速度和效率，LightGBM 在許多資料科學家和機器學習工程師中成為了首選工具，特別適用於需要快速處理龐大數據量的情境，例如在廣告排序、網絡流量預測等領域。無論是獨立使用或與其他模型結合，LightGBM 都能提供卓越的預測效能。

3.3 模型評估方法

繼上一節詳細介紹了四種機器學習演算法後，我們接下來探討如何判斷這些演算法的優劣。合適的評估標準不僅能精準地量化模型的預測能力，也能驗證其在實際應用中的效果。因此，本節重點介紹本研究採用的評估工具，包括混淆矩陣、ROC 和 AUC，以及運算時間，這有助於我們全面理解所採用模型的性能及其實用價值。

3.3.1 混淆矩陣

混淆矩陣（Confusion Matrix）是評估分類模型性能的一個關鍵工具，它透過表格形式展現了模型預測的結果與實際情況的對比。二類別的混淆矩陣如表 1 所示，我們將藉此計算下列四種重要指標，從不同角度解釋模型的預測效能：

- 精確率（Precision）：預測為正類別的樣本中實際為正類別的比例，即 TP / (TP + FP)。
- 召回率（Recall）：實際為正類別的樣本中預測為正類別的比例，即 TP / (TP + FN)。
- 準確率（Accuracy）：預測正確的樣本數佔所有樣本數的比例，即 (TP + TN) / (TP + TN + FP + FN)。
- F1 分數（F1 Score）：精確率與召回率的調和平均，用來綜合評估模型在正類別的表現，其計算公式為 $2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$。

由本研究著重購買行為預測並應用在行銷計劃的執行，精確率有利於我們評估行銷成本。精確率是由真陽性除以所有預測為正例的總數（真陽性加假陽性）計算得出，它衡量的是被模型預測為正例的樣本中，實際為正例的比例。在行銷情境中，高精確率意味著較少的資源會被浪費在錯誤識別的客戶上，從而提高行銷活動的效益。

雖然具體的可接受精確率標準會根據具體的行銷策略和目標而有所不同，但一般來說，在行銷應用中，如果精確率能達到或超過 0.7，則通常被認為是可接受的。這意

表 1：混淆矩陣。

混淆矩陣		真實	
		正	負
預測	正	真陽性 （True Positive）	假陽性 （False Positiv）
	負	假陰性 （False Negative）	真陰性 （True Negative）

味著至少 70% 被模型識別為具有購買潛力的潛在客戶實際上符合這一預測，這樣的效率對於大多數行銷活動來說是有效的。然而，根據不同的行銷成本結構和預期回報，某些情況下可能需要更高的精確率。

3.3.2 ROC 和 AUC

ROC 曲線（Receiver Operating Characteristic Curve）是通過在不同閾值下計算真陽性率（TPR）和假陽性率（FPR）所繪製出的圖形，範例如圖 4 所示。這一曲線描繪了分類器在各種靈敏度水平上的表現。

所謂「閾值」是指在分類模型預測結果為正類或負類時所設定的判定標準。不同的閾值會影響模型對預測結果劃分為正類或負類的傾向，進而影響模型性能的評估。

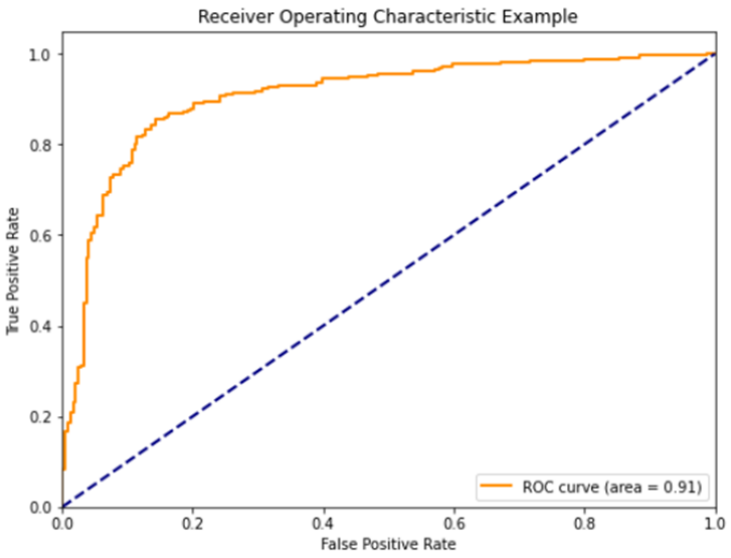

圖 4：ROC 曲線示意圖。

進一步地，真陽性率（TPR），也被稱為敏感度，指的是在所有實際為正類的樣本中，模型正確識別出的比例，計算公式為真陽性（TP）除以真陽性與假陰性（FN）的總和。而假陽性率（FPR）則是指在所有實際為負類的樣本中，模型錯誤地識別為正類的比例，計算公式為假陽性（FP）除以假陽性與真陰性（TN）的總和。

ROC 曲線下的面積，即 AUC（Area Under the Curve），是用來綜合評估模型優劣的一個指標。AUC 越接近 1，表示模型的分類性能越好；相反，AUC 越接近 0.5，則意味著模型的性能僅等同於隨機猜測。針對商業領域的行銷場景，通常 AUC 在 0.7 以上即為合適可接受的效能。

3.3.3 運算花費時間

我們在選擇模型時，除了評估效能指標外，運算花費時間也是一個重要的考慮因素。尤其是在需要處理大規模數據集或透過雲端服務運算的場景中，運算效率成為影響模型選擇的關鍵。模型的運算效率直接影響其部署和運營成本，一個高效能但計算需求巨大的模型，在擁有有限計算資源的環境中可能不實用。在商業應用中，計算時間長會導致更高的能源消耗和成本支出，這可能影響整體投資回報率。

總結而言，運算花費時間作為一項輔助評估標準，幫助我們在效能、成本和實際應用需求之間取得平衡，確保選擇的模型不僅在技術層面上優異，同時也在實際操作環境中具備可行性與經濟效益。

4. 研究結果

本研究旨在利用客戶消費記錄來預測未來的購買行為，並區分客戶為一般客戶或高貢獻客戶。資料來自 H&M 的商品資訊、客戶資訊及交易記錄等共計 32 個變數，希望可以截取資料的特徵，以預測未來 1 個月及 3 個月的購買行為或客戶貢獻度，期望可應用於短期促銷及季度行銷規劃。

為了讓資料結構呈現更清楚，敘述統計先挑選 2 個時間區間的資料進行分析，模型建立則以從過往資料挑選不同區間的資料，分別以 4 種機器學習演算法對未來 1 個月和 3 個月進行預測。評估模型時，除了考慮 AUC 分數，有別於普遍採用 F1-score，本研究側重行銷策劃的成效，以精確率（precision）比較不同模型及區間的表現。最後，針對成效符合預期的模型，檢視重要影響變數。

4.1 資料來源及預處理

4.1.1 資料來源說明

本研究資料來源為 2022 年 H&M 於 Kaggle 發布的競賽：「H&M Personalized Fashion Recommendations」（https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations），該競賽所提供之數據包含商品資訊、客戶資訊以及交易記錄等內容。其中，商品資訊共包含 25 個欄位，總筆數為 105,542；我們以 article_id 的欄位與交易記錄進行串接。又客戶資訊共有 7 個欄位，總筆數為 1,371,980，以 customer_id 欄位與交易記錄進行串接。最後，交易記錄共有 5 個欄位，總筆數為 31,788,324，線上佔 70.40%、門市佔 29.60%，區間為 2018 年 9 月至

表 2：商品資訊。

編號	欄位名稱	欄位說明	範例資料
1	article_id	商品 ID	0671433001
2	product_code	商品代碼	0671433
3	prod_name	商品名稱	ED Dottie blouse
4	detail_desc	詳細描述	Blouse in an airy weave with a small stand-up...
5	index_group_no	主分類代碼	1
6	index_group_name	主分類名稱	Ladieswear
7	index_code	次分類代碼	A
8	index_name	次分類名稱	Ladieswear
9	product_group_name	商品主類型名稱	Garment Upper body
10	product_type_no	商品次類型代碼	258
11	product_type_name	商品次類型名稱	Blouse
12	garment_group_no	服裝類型代碼	1010
13	garment_group_name	服裝類型名稱	Blouses
14	section_no	商品主系列代碼	2
15	section_name	商品主系列名稱	H&M+
16	department_no	商品次系列代碼	1909
17	department_name	商品次系列名稱	Woven top
18	graphical_appearance_no	外觀樣式代碼	1010001
19	graphical_appearance_name	外觀樣式名稱	All over pattern
20	perceived_colour_value_id	色調代碼	4
21	perceived_colour_value_name	色調名稱	Dark
22	perceived_colour_master_id	主要顏色代碼	2
23	perceived_colour_master_name	主要顏色名稱	Blue
24	colour_group_code	顏色組合代碼	73
25	colour_group_name	顏色組合名稱	Dark Blue

表 3：客戶資訊。

編號	欄位名稱	欄位說明	範例資料
1	customer_id	客戶 ID	ac2421209ada389f4a89...
2	postal_code	郵政編碼	ff78946468e6db4112a5...
3	age	年齡	48
4	club_member_status	會員狀態	ACTIVE
5	FN	接受廣宣	1
6	Active	對廣宣活躍	1
7	fashion_news_frequency	廣宣頻率	Regularly

表 4：交易記錄。

編號	欄位名稱	欄位說明	範例
1	customer_id	客戶 ID	000058a12d5b43e67d...
2	article_id	商品 ID	0663713001
3	sales_channel_id	通路 ID	2
4	price	銷售金額	0.050831
5	date	銷售日期	2018-09-20

2020 年 9 月，欄位說明及範例資料分別列於表 2、表 3、表 4。三項資訊整合及預處理後，於 1 個月內有消費的客戶筆數約 7 萬，全部期間消費的客戶約 86 萬筆。

4.1.2 資料預處理

本研究旨在利用過往行為記錄，即整合商品資訊、客戶資訊及交易記錄，來預測未來客戶的購買行為，並判斷客戶為一般客戶或高貢獻客戶，將「是否購買」與「貢獻分類」作為預測標籤（Labels）。此處高貢獻客戶的定義參考帕雷托法則，即前 20% 的客戶可為企業帶來 80% 的收益，因此本研究以累計消費金額達前 20 百分位以上的消費客戶歸類為高貢獻客戶。

我們設定預測區間為未來 1 個月及 3 個月，期望可以分別適用於短期促銷活動及下一季度新品宣傳等兩種行銷情境。而過往交易記錄的時間範圍，則選擇共 7 個時間區間：過去 1 個月、2 個月、3 個月、6 個月、9 個月、12 個月及全部資料。

我們分別計算 2 個預測區間的預測標籤，以及利用 7 個時間區間進行特徵工程（Feature Engineering），因此共有 14 種時間組合。針對這些組合以代碼標示，如「以過往 1 個月資料預測未來 1 個月行為」標示為「X1y1」，又「以全部歷史資料預測未來 3 個月行為」標示為「XHy3」，以此類推。接下來訂定模型的訓練期（Training

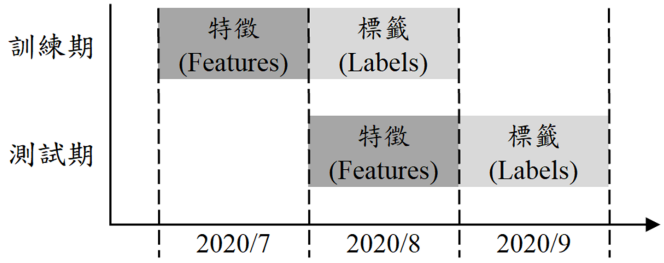

圖 5：X1y1 訓練期及測試期區間示意圖。

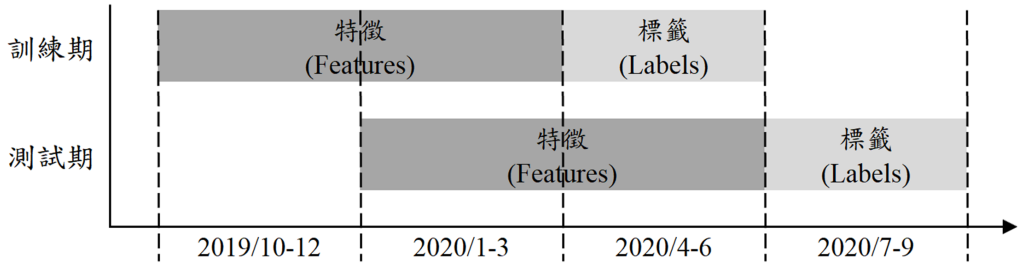

圖 6：X6y3 訓練期及測試期區間示意圖。

表 5：模型區間代碼時間對照表。

區間代碼	訓練期（Training Period）		測試期（Test Period）	
	特徵（Features）計算區間	標籤（Labels）計算區間	特徵（Features）計算區間	標籤（Labels）計算區間
X1y1	2020/7		2020/8	
X2y1	2020/6-2020/7		2020/7-2020/8	
X3y1	2020/5-2020/7		2020/6-2020/8	
X6y1	2020/2-2020/7	2020/8	2020/3-2020/8	2020/9
X9y1	2019/11-2020/7		2019/12-2020/8	
X12y1	2019/8-2020/7		2019/9-2020/8	
XHy1	2018/9-2020/7		2018/9-2020/8	
X1y3	2020/3		2020/6	
X2y3	2020/2-2020/3		2020/5-2020/6	
X3y3	2020/1-2020/3		2020/4-2020/6	
X6y3	2019/10-2020/3	2020/4-2020/6	2020/1-2020/6	2020/7-2020/9
X9y3	2019/7-2020/3		2019/10-2020/6	
X12y3	2019/4-2020/3		2019/7-2020/6	
XHy3	2018/9-2020/3		2018/9-2020/6	

表 6：預測標籤清單。

編號	預測標籤（Labels）	說明	範例資料
1	y_d1	是否購買： 0 = 無購買 1 = 有購買	1
2	y_d2	貢獻分類： 0 = 無購買 1 = 一般客戶 2 = 高貢獻	2

Period）及測試期（Test Period），其各自的預測標籤（Labels）之計算區間必須完全分隔，可透過圖 5 及圖 6 舉例說明。

圖 5 為「X1y1」的訓練期與測試期示意圖，訓練期的特徵計算期間為 2020 年 7 月，標籤計算期間為 2020 年 8 月；測試期的特徵計算期間為 2020 年 8 月，標籤計算期間為 2020 年 9 月。而如圖 6 則為「X6y3」的示意圖，訓練期的特徵為 2019 年 10 月至 2020 年 3 月，標籤為 2020 年 4 月至 6 月；測試期的特徵為 2020 年 1 月至 6 月，標籤為 2020 年 7 月至 9 月。完整 14 項區間的時間對照表可參考表 5。

表 6 為預測標籤清單，我們分別定義兩項預測目標如下：「y_d1」代表客戶未來是否購買、「y_d2」則代表客戶的貢獻分類。

經客戶資訊初步探索後，客戶會員狀態為「ACTIVE」佔全體客戶 93.16%，年齡資料無缺失值者佔 98.84%，因此我們篩選客戶會員狀態為「ACTIVE」，且具有年齡資料者進行後續分析，總筆數為 1,266,255。此外，交易記錄依表 5 的日期區間篩選，考量僅購買 1 次的客戶可供分析的行為過少，故排除特徵計算期間僅購買 1 次的客戶。彙整篩選後客戶資訊、交易記錄及商品資訊，總共 32 項特徵如表 7 所列，除了年齡以外，其餘變數皆經過加工處理，以下將逐一說明。

首先說明表 7 編號 2-4 郵政頻率相關的變數，此為觀察原始郵政編碼的次數分佈後，發現其中一組編碼將近 12 萬次，與次數排名第二的 260 次相差懸殊，如表 8。由於 H&M 未於 Kaggle 對此多做說明，推測此編碼可能為特殊的地點，抑或是由空值加密而成，因此新增一欄郵政頻率（postal_freq），註記此編碼為「special」；排除此編碼後，將其餘編碼的出現次數以第三四分位數加上 1.5 倍 IQR 作為離群值，對次數高於離群值的編碼註記為「outlier」，其餘則註記為「normal」，最後轉換為虛擬變數共 3 欄。

而在客戶的廣宣意願方面，原始資料共有 3 個欄位：接受廣宣、對廣宣活躍、廣

表 7：特徵變數清單。

編號	類型	特徵變數	說明	範例資料
1	連續	age	年齡	47
2	類別	postal_freq_normal	郵政頻率 _ 一般	0, 1
3	類別	postal_freq_outlier	郵政頻率 _ 高頻	0, 1
4	類別	postal_freq_special	郵政頻率 _ 特殊	0, 1
5	類別	news_active_flg	廣宣頻繁活躍	0, 1
6	連續	quantity	購買數量	15
7	連續	baby/children	購買數量 _ 商品 _ 小孩	1
8	連續	divided	購買數量 _ 商品 _ 特殊品	0
9	連續	ladieswear	購買數量 _ 商品 _ 女性	14
10	連續	menswear	購買數量 _ 商品 _ 男性	0
11	連續	sport	購買數量 _ 商品 _ 運動	0
12	連續	popular	購買數量 _ 熱銷品	8
13	連續	weekend	購買數量 _ 週末	10
14	連續	period	第 1 筆及最後 1 筆間距天數	49
15	連續	first_buy_days	歷史第 1 筆與特徵最後日間距天數	549
16	連續	discount	平均折扣	0.485745
17	連續	r	Recency 最近一次消費	0.216017
18	連續	f	Frequency 消費次數	3
19	連續	m	Monetary 消費金額	45
20	類別	rfmcluster	RFM 分群	0, 1, 2, 3
21	連續	online_period	線上 _ 第 1 筆及最後 1 筆間距天數	0
22	連續	store_period	門市 _ 第 1 筆及最後 1 筆間距天數	49
23	連續	online_discount	線上 _ 平均折扣	0.665215
24	連續	store_discount	門市 _ 平均折扣	0.458135
25	連續	online_r	線上 _R	0.049119
26	連續	online_f	線上 _F	1
27	連續	online_m	線上 _M	61
28	連續	store_r	門市 _R	0.166898
29	連續	store_f	門市 _F	2
30	連續	store_m	門市 _M	33
31	類別	online_rfmcluster	線上 _RFM 分群	0, 1, 2, 3
32	類別	store_rfmcluster	門市 _RFM 分群	0, 1, 2, 3

表 8：郵政編碼次數分配表（前 3 名）。

郵政編碼	次數
2c29ae653a9282cce4151bd87643c907644e09541abc28ae87dea0d1f6603b1c	117817
cc4ed85e30f4977dae47662ddc468cd2eec11472de6fac5ec985080fd92243c8	260
714976379549eb90aae4a71bca6c7402cc646ae7c40f6c1cb91d4b5a18623fc1	158

表 9：廣宣意願次數分配表。

接受廣宣	對廣宣活躍	廣宣頻率	客戶 ID 次數	客戶 ID 佔比
null	null	None	892931	65.08%
		Monthly	13	0.00%
		Regularly	2106	0.15%
1	null	None	291	0.02%
		Monthly	31	0.00%
		Regularly	12204	0.89%
	1	None	500	0.04%
		Monthly	798	0.06%
		Regularly	463106	33.75%

宣頻率。依此分類可知其各自的客戶分佈如表 9，可以進一步將客戶分為兩類，第一類底色為灰色者（即為表 9 最後兩列），他們接受廣宣、對廣宣活躍且廣宣頻率為「Regularly」及「Monthly」，歸為廣宣頻繁活躍者；第二類為無底色者，他們不對廣宣活躍或沒有廣宣頻率，則歸為對廣宣無註記或無反應。記錄此分類結果於新增的廣宣頻繁活躍（news_active_flg）欄位中，第一類註記為 1，第二類則註記為 0，如表 7 變數編號 5。

因原始交易記錄無商品數量，故將每一個商品 ID 皆視為售出 1 件，表 7 編號 6 的購買數量，即為特徵計算期間的數量總和。而編號 7-11 則為根據商品資訊的主分類名稱（index_group_name），計算出各品類購買數量。若某一商品在特徵計算期間內總銷量達前 20 百分位，則註記其為熱銷品，而客戶於期間內購買的熱銷品數量則記錄於編號 12。而考量週間與週末可能有不同的消費行為，記錄編號 13 客戶為於週末購買的商品數量。透過表 7 的範例資料來說明，編號 6 購買數量為 15 件，其中 8 件為熱銷品，而有 10 件商品於週末時購入。

此外，我們認為客戶接觸品牌的時間亦有可能影響消費行為，編號 14 為記錄客戶於特徵計算期間第 1 筆及最後 1 筆消費的間距天數。由於客戶資訊中並無客戶首次消費或是加入會員的時間，因此我們視全部資料中第 1 筆消費為首購日，並計算與特徵計算期間迄日的間距天數，此為編號 15。

因在商品資訊中並無商品的定價，為計算平均折扣（編號 16），將特徵計算期間該商品的最高售價視作定價，進而算出每項商品的折扣，而客戶購買的平均折扣即為其購買的商品折扣平均值。

編號 17-19 分別為 RFM 指標，R 為最近一次消費（Recency），是每個客戶於特徵計算期間內最後 1 筆消費日期，至特徵計算期間迄日的間距天數。F 為消費次數（Frequency），此處計算客戶於特徵計算期間的消費次數，同一天購買多個商品判斷為一次消費。M 為消費金額（Monetary），計算客戶於特徵計算期間的總消費金額。編號 20 則是透過 K-Means 方法，對 RFM 這 3 個變數同時進行分群，根據第三章提到的分群選擇策略，以 4 群作為分群數量，並依照各群的平均消費金額，由低至高給予分群編號 0 到 3。

最後，由於線上及門市通路可能存在不同消費行為，將 6 項特徵依通路進行細分：第 1 筆及最後 1 筆間距天數、平均折扣、R、F、M 及 RFM 分群等，細分後的特徵為編號 21-32。除了 RFM 分群之外，其他 5 項特徵區分通路計算時，對於無消費者的資料給予特定數值，包含平均折扣記為 -1，第 1 筆及最後 1 筆間距天數、最近一次消費記為 -100，消費次數、消費金額（M）記為 0。這些特定數值的設置，是為了保留無消費記錄的客戶，避免數據遺失。在分析過程中亦測試了其他數值，發現這些數值對分析結果影響不大，故選擇此組特定數值進行統一處理。以上變數，除了編號 1-5 屬於客戶基本資料，其他編號 6-32 皆萃取自交易記錄，因此這些變數會隨著特徵計算區間跟著有所變動。

4.2 敘述統計

本研究的敘述統計分為三個部分，分別對應變數－預測標籤、類別型特徵變數及連續型特徵變數分析。考量到本研究共有 14 種時間區間的資料，因此從中挑選「以過往 3 個月資料預測未來 3 個月行為」（X3y3）和「以過往 9 個月資料預測未來 3 個月行為」（X9y3），於帶入模型訓練前，以訓練期的資料比較兩區間資料結構的差異。

4.2.1 應變數－預測標籤分析

表 10 記錄 X3y3 與 X9y3 分別在是否購買（y_d1）和貢獻分類（y_d2）的分佈狀況。首先觀察有購買者（y_d1 = 1），X3y3 相比 X9y3 高出了 14.24%；接著是貢獻分類（y_d2），由於高貢獻客戶（y_d2 = 2）是來自有購買者（y_d1 = 1）中消費前 20% 客戶，故高貢獻客戶（y_d2 = 2）在 X3y3 與 X9y3 的佔比相同，而一般客戶（y_d2 = 1）同為 X3y3 佔比高於 X9y3。我們猜測原因為較近期購買的客戶，他們對於品牌比較有印象，如果對先前的購物體驗感受良好，那麼一旦有新需求出現，就有可能再次回購同一品牌。

表 10：應變數百分比次數分配表。

區間代碼	總筆數	是否購買（y_d1）			貢獻分類（y_d2）			
		0	1	合計	0	1	2	合計
X3y3	209,564	25.30%	74.70%	100.00%	25.30%	54.70%	20.00%	100.00%
X9y3	529,993	39.54%	60.46%	100.00%	39.54%	40.46%	20.00%	100.00%

表 11：兩區間類別型特徵變數與預測標籤比較。

特徵	區間代碼	類別	是否購買（y_d1）		貢獻分類（y_d2）		
			0	1	0	1	2
郵政頻率	X3y3	一般	77.99%	79.40%	77.99%	79.49%	79.17%
		高頻	18.04%	19.65%	18.04%	19.33%	20.53%
		特殊	3.96%	0.94%	3.96%	1.18%	0.30%
	X9y3	一般	77.29%	79.45%	77.29%	79.52%	79.30%
		高頻	17.85%	19.40%	17.85%	18.95%	20.30%
		特殊	4.86%	1.16%	4.86%	1.53%	0.40%
廣宣頻繁活躍	X3y3	是	37.01%	46.81%	37.01%	46.22%	48.43%
		否	62.99%	53.19%	62.99%	53.78%	51.57%
	X9y3	是	37.61%	44.62%	37.61%	44.47%	44.92%
		否	62.39%	55.38%	62.39%	55.53%	55.08%
RFM 分群	X3y3	0	18.54%	10.34%	18.54%	11.32%	7.66%
		1	25.79%	20.02%	25.97%	21.35%	16.36%
		2	33.90%	34.82%	33.90%	35.61%	32.67%
		3	21.78%	34.82%	21.78%	31.72%	43.31%
	X9y3	0	16.13%	5.99%	16.13%	6.99%	3.96%
		1	26.96%	16.47%	26.96%	18.56%	12.25%
		2	27.41%	26.50%	27.41%	28.02%	23.42%
		3	29.49%	51.04%	29.49%	46.43%	60.37%

4.2.2 類別型特徵變數分析

基於上述結果，我們進一步分別比較是否購買（y_d1）及貢獻分類（y_d2），他們在其他類別型變數的分佈，並彙整結果於表 11。首先為郵政頻率，表 11 顯示 X3y3 和 X9y3 兩組資料分佈相近，值得注意的類別是郵政頻率「特殊」，在是否購買中，無購買者（y_d1 = 0）的「特殊」佔比高於有購買者（y_d1 = 1）；同樣在貢獻分類

表 12：兩區間類別型特徵變數與預測標籤比較－特定類別。

特徵	區間代碼	類別	是否購買（y_d1）		貢獻分類（y_d2）		
			0	1	0	1	2
郵政頻率	X3y3	特殊	58.76%	41.24%	58.76%	37.74%	3.49%
	X9y3	特殊	73.32%	26.68%	73.32%	23.62%	3.06%
廣宣頻繁活躍	X3y3	是	21.12%	78.88%	21.12%	57.03%	21.85%
	X9y3	是	35.54%	64.46%	35.54%	42.99%	21.47%
RFM 分群	X3y3	3	17.48%	82.52%	17.48%	55.04%	27.48%
	X9y3	3	27.43%	72.57%	27.43%	44.18%	28.39%

（y_d2）上，無消費者（y_d2 = 0）的「特殊」佔比大於一般客戶（y_d2 = 1），又一般客戶佔比大於高貢獻客戶（y_d2 = 2）。於是我們進一步想知道在郵政頻率為「特殊」之下，是否購買（y_d1）及貢獻分類（y_d2）的分佈，原本應該列出所有類別下是否購買（y_d1）及貢獻分類（y_d2）的分佈數據，但礙於篇幅我們僅挑選特定類別記錄於表 12。從表 12 我們可以看到 X3y3 資料在郵政頻率為「特殊」的類別下，無購買者為 58.76% 高於有購買者 41.24%，在貢獻分類當中無購買者仍為 58.76%、一般客戶為 37.74%、高貢獻客戶 3.49%。X9y3 的形況之下，無購買者、一般客戶和高貢獻客戶則分別為 73.32%、23.62% 和 3.06%。我們觀察到無論 X3y3 或 X9y3，在郵政頻率為「特殊」之下，無購買者的佔比都高於有購買者。

第二項為廣宣頻繁活躍，X3y3 和 X9y3 兩組資料分佈相近，有購買者（y_d1 = 1）中廣宣頻繁活躍者的佔比高於無購買者（y_d1 = 0），而高貢獻客戶（y_d2 = 2）與一般客戶（y_d2 = 1）的廣宣頻繁活躍者的佔比相近。接著深入分析廣宣頻繁活躍者，由表 12 可觀察 X3y3 中的資料，無購買者、一般客戶與高貢獻客戶的佔比分別為 21.12%、57.03% 與 21.85%，X9y3 則分別為 35.54%、42.99% 與 21.47%。由此可知對於廣宣頻繁活躍者，較有機會成為未來有購買者，且在特徵計算期間較短時更加明顯。

接下來為 RFM 分群，分群編號 0 皆為 R 大、F 及 M 小，分群編號 3 為 R 小、F 及 M 大，依序代表為許久未購買且貢獻度最低、稍久未購買且貢獻度低、前陣子有購買且貢獻度中、近期購買且貢獻度高。從表 11 可看出 X3y3 內有購買客戶（y_d1 = 1）的 RFM 分群編號 3 與編號 2 並列佔比最多者，同時無購買者（y_d1 = 0）分群編號 2 佔多數；在貢獻分類中，高貢獻客戶（y_d2 = 2）其 RFM 分群編號 3 最多，一般客戶（y_d2 = 1）則是分群編號 2 為最多。而在 X9y3 區間，

無論是否購買，佔比均隨著分群編號遞增，其中有購買者以編號 3 佔比最多，且約為無購買者的兩倍。其次，比較貢獻分類（y_d2）的分佈狀況，分群編號 3 的佔比差距隨著貢獻分類增加。我們進一步透過表 12 檢視分群編號 3 底下，X3y3 資料中無購買者、一般客戶與高貢獻客戶的佔比分別為 17.48%、55.04% 與 27.48%，X9y3 資料中則分別為 27.43%、44.18% 與 28.39%，如此看來 RFM 分群編號 3 的客戶更有機會成為未來有購買的客戶。

4.2.3 連續型特徵變數分析

我們轉向分析連續型變數在應變數間的平均值差異，惟表 7 特徵變數清單的編號 21-30，由於在區分線上或門市通路時，已針對無消費者的資料給予特定數值，詳如上一節資料預處理最後一段說明，故這些變數不納入討論。我們記錄連續型變數與是否購買（y_d1）的平均值比較於表 13，與貢獻分類（y_d2）的平均值比較記錄於表 14。事實上，我們亦對表 12 中的平均值進行 t 檢定，對表 13 的平均值進行 ANOVA 檢定，所有的檢定結果皆為顯著。

首先，變數編號 1 為年齡，兩組資料是在是否購買（y_d1）跟貢獻分類（y_d2）平均值似乎無明顯不同，不過因為樣本數很大，因此檢定結果仍然顯著。接著為購買數量相關特徵（編號 6-13），多數特徵有明顯差異，有購買者（y_d1 = 1）的平均購買數量高於無購買者（y_d1 = 0），而高貢獻客戶（y_d2 = 2）又高於與一般客戶（y_d2 = 1）。不過商品為小孩、男性及運動這三類品項，其銷售數量相比女性及特殊品項較少，因而差異較小。

接著為編號 14 的第 1 筆及最後 1 筆間距天數，我們能發現有購買者（y_d1 = 1）或高貢獻客戶（y_d2 = 2），其平均值皆較高。我們猜測間距天數較長的客戶，可能於期間內持續關注品牌，因此未來再次消費機會高。編號 15 為歷史第 1 筆與特徵最後日間距天數，也可以明顯看出，有購買者（y_d1 = 1）或高貢獻客戶（y_d2 = 2），其平均值皆較高，因此可以說明客戶接觸品牌的時間越長，未來越可能持續消費。而編號 16 平均折扣，對於是否購買（y_d1）跟貢獻分類（y_d2）其平均值差異不大。

最後為 RFM 指標（編號 17-19），有購買者（y_d1 = 1）或高貢獻客戶（y_d2 = 2）的最近一次消費（R）的平均值較小，表示最近購買的客戶更有可能未來再次購買，甚至成為高貢獻客戶。而消費次數（F）和消費金額（M）的狀況相似，兩指標平均值皆隨著貢獻度增高，尤其是消費金額（M）尤其明顯，其高貢獻客戶的平均值高於一般客戶兩倍以上。因此，我們可以推論消費次數或消費金額越高的客戶，未來回購機會

表 13：兩區間連續型特徵變數與是否購買比較。

編號	特徵	X3y3		X9y3	
		0	1	0	1
1	年齡	36.50	35.76	36.55	35.58
6	購買數量	8.22	11.37	12.34	22.21
7	購買數量 _ 商品 _ 小孩	0.16	0.27	0.28	0.64
8	購買數量 _ 商品 _ 特殊品	1.93	2.64	2.94	5.16
9	購買數量 _ 商品 _ 女性	5.24	7.32	7.85	14.23
10	購買數量 _ 商品 _ 男性	0.41	0.53	0.74	1.27
11	購買數量 _ 商品 _ 運動	0.47	0.61	0.53	0.90
12	購買數量 _ 熱銷品	6.96	9.58	10.39	18.47
13	購買數量 _ 週末	2.42	3.33	3.54	6.40
14	第 1 筆及最後 1 筆間距天數	29.02	40.51	115.98	164.15
15	歷史第 1 筆與特徵最後日間距天數	379.77	460.47	391.79	451.83
16	平均折扣	0.87	0.86	0.84	0.83
17	Recency 最近一次消費	33.43	26.11	89.98	57.82
18	Frequency 消費次數	2.52	3.41	3.69	6.58
19	Monetary 消費金額	0.23	0.32	0.34	0.63

表 14：兩區間連續型特徵變數與貢獻分類比較。

編號	特徵	X3y3			X9y3		
		0	1	2	0	1	2
1	年齡	36.50	35.64	36.10	36.55	35.44	35.85
6	購買數量	8.22	9.03	17.78	12.34	16.11	34.56
7	購買數量 _ 商品 _ 小孩	0.16	0.20	0.46	0.28	0.45	1.04
8	購買數量 _ 商品 _ 特殊品	1.93	2.13	4.02	2.94	3.84	7.84
9	購買數量 _ 商品 _ 女性	5.24	5.74	11.65	7.85	10.16	22.48
10	購買數量 _ 商品 _ 男性	0.41	0.46	0.71	0.74	1.01	1.79
11	購買數量 _ 商品 _ 運動	0.47	0.49	0.95	0.53	0.65	1.41
12	購買數量 _ 熱銷品	6.96	7.64	14.87	10.39	13.52	28.49
13	購買數量 _ 週末	2.42	2.67	5.16	3.54	4.64	9.95
14	第 1 筆及最後 1 筆間距天數	29.02	38.22	46.79	115.98	154.96	182.75
15	歷史第 1 筆與特徵最後日間距天數	379.77	449.33	490.93	391.79	439.82	476.11
16	平均折扣	0.87	0.85	0.87	0.84	0.83	0.84
17	Recency 最近一次消費	33.43	27.40	22.60	89.98	62.95	47.44
18	Frequency 消費次數	2.52	3.08	4.31	3.69	5.51	8.74
19	Monetary 消費金額	0.23	0.24	0.54	0.34	0.43	1.04

也較高。

4.3 模型預測

本研究針對客戶的購買行為以及貢獻度進行預測，使用邏輯斯迴歸、決策樹、XGBoost 及 LightGBM 等 4 種機器學習演算法。我們初期設定了兩個預測目標：是否購買（y_d1）以及貢獻分類（y_d2），並使用了 14 種區間進行測試，區間代碼及其相對應的時間範圍詳如本章表 5。不過，以全部歷史數據作為訓練期資料並對客戶未來 1 個月的貢獻度分類時，由於未來 1 個月有購買記錄的客戶僅佔全部歷史數據的 19%，因此無法再將客戶進一步分為一般與高貢獻客戶（前 20%），故「以全部歷史資料預測未來 1 個月行為」（XHy1），此區間不進行貢獻度預測。

為了綜合評估模型成效，我們採用 AUC 分數與精確率（precision）這兩項普遍認可的指標，並認定兩者均達到 0.7 表示模型具有良好預測能力。然而企業制定行銷計畫時，不僅要考慮預測準確度，還需謹慎平衡目標營收和行銷預算之間的效益。換言之，預測購買人數乘上單位行銷成本應該低於整體行銷預算，而實際購買人數乘上平均客單價則應超越訂定的目標營收，以此來確保行銷活動投資報酬率。基於這些考量，我們將 AUC 分數、精確率、召回率、預測購買人數及實際購買人數列入表格，從而比較不同演算法和資料區間的成效。

就預測成效而言，雖然大多數演算法對於預測客戶是否購買，都有一些表現良好的區間，但在區分一般客戶和高貢獻客戶上，結果普遍不盡如人意。這可能是由於客戶被分為三個群體後，各群之間的差異性減弱所致。於是，我們新增設了第三個預測目標，聚焦預測客戶未來是否會成為高貢獻客戶，儘管調整目標後結果略見提升，但所有演算法的 AUC 分數與精確率皆未達 0.7 以上。

4.3.1 邏輯斯迴歸

藉由邏輯斯迴歸預測是否購買、貢獻分類和是否為高貢獻客戶，結果分別記錄於表 15、表 16 和表 17，運算時間依序為 4 分 11 秒、5 分 59 秒和 4 分 11 秒。

在是否購買的預測中，表 15 顯示預測未來 1 個月（X1y1∼XHy1）的區間，AUC 最大值為 0.76，隨著使用的過往資料越多同步遞增，然而精確率均落在 0.5 上下，對於購買者的預測表現差。在預測未來 3 個月（X1y3∼XHy3）的區間，AUC 最大值為 0.78，同樣隨資料增加而遞增，而精確率最大值為 0.80，卻是隨著使用過往資料增加而遞減，兩者均大於 0.7 的區間為 X2y3、X3y3 和 X6y3。從行銷策略來看，若期望行

表 15：邏輯斯迴歸－是否購買各區間結果比較表。

區間代碼	AUC	未來會購買			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.65	0.52	0.76	26,345	50,834
X2y1	0.65	0.51	0.52	35,950	69,944
X3y1	0.66	0.49	0.49	46,645	94,625
X6y1	0.69	0.51	0.40	49,250	95,939
X9y1	0.71*	0.53	0.34	47,766	90,458
X12y1	0.72*	0.53	0.33	49,065	93,375
XHy1	0.76*	0.52	0.30	47,912	92,412
X1y3	0.69	0.8*	0.99	80,653	100,995
X2y3	0.7*	0.77*	0.98	140,804	183,520
X3y3	0.72*	0.74*	0.97	176,270	239,137
X6y3	0.72*	0.7*	0.92	245,575	352,264
X9y3	0.74*	0.68	0.84	268,607	395,860
X12y3	0.75*	0.68	0.76	267,218	391,214
XHy3	0.78*	0.68	0.71	271,347	400,311

* 表 AUC 或精確率達到 0.7

表 16：邏輯斯迴歸－貢獻分類各區間結果比較表。

區間代碼	AUC	一般客戶				高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數	精確率	召回率	實際購買人數	預測購買人數
X1y1	0.65	0.35	0.29	5,608	16,149	0.47	0.27	4,161	8,772
X2y1	0.65	0.32	0.14	4,603	14,565	0.50	0.22	7,927	15,967
X3y1	0.66	0.27	0.13	5,126	19,014	0.50	0.22	12,374	24,849
X6y1	0.68	0.24	0.05	2,008	8,498	0.51	0.24	20,501	40,565
X9y1	0.7*	0.22	0.03	792	3,656	0.53	0.23	24,498	46,476
X12y1	0.72*	0.20	0.01	275	1,396	0.54	0.24	29,555	54,932
X1y3	0.69	0.62	0.94	57,060	91,980	0.57	0.25	5,070	8,832
X2y3	0.7*	0.59	0.90	95,117	160,652	0.57	0.31	11,924	21,073
X3y3	0.71*	0.56	0.86	113,237	200,979	0.55	0.35	18,196	32,882
X6y3	0.72*	0.52	0.75	139,535	266,824	0.58	0.35	29,330	50,658
X9y3	0.72*	0.49	0.62	130,773	265,122	0.60	0.35	37,497	62,492
X12y3	0.73*	0.48	0.51	112,359	232,294	0.62	0.33	41,682	67,207
XHy3	0.75*	0.45	0.37	79,871	179,048	0.59	0.38	61,980	105,111

* 表 AUC 或精確率達到 0.7

表 17：邏輯斯迴歸－是否為高貢獻客戶各區間結果比較表。

區間代碼	AUC	高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.7*	0.57	0.16	2,378	4,144
X2y1	0.69	0.58	0.13	4,481	7,673
X3y1	0.69	0.58	0.13	7,276	12,643
X6y1	0.7*	0.58	0.15	12,617	21,653
X9y1	0.71*	0.60	0.15	15,766	26,285
X12y1	0.72*	0.60	0.16	19,599	32,504
XHy1	0.76*	0.60	0.19	29,746	49,568
X1y3	0.75*	0.59	0.23	4,839	8,251
X2y3	0.76*	0.59	0.29	10,986	18,735
X3y3	0.77*	0.59	0.31	16,081	27,230
X6y3	0.77*	0.64	0.29	24,325	38,132
X9y3	0.78*	0.66	0.28	30,317	45,772
X12y3	0.78*	0.69	0.26	33,176	48,344
XHy3	0.79*	0.65	0.30	49,533	75,639

* 表 AUC 或精確率達到 0.7

表 18：決策樹－是否購買各區間結果比較表。

區間代碼	AUC	未來會購買			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.54	0.49	0.60	20,906	42,871
X2y1	0.54	0.43	0.54	37,121	87,105
X3y1	0.55	0.39	0.52	49,747	128,018
X6y1	0.56	0.35	0.48	60,012	169,350
X9y1	0.57	0.33	0.45	63,305	191,957
X12y1	0.58	0.32	0.44	65,069	200,952
XHy1	0.59	0.30	0.40	63,094	213,620
X1y3	0.54	0.8*	0.83	67,434	83,781
X2y3	0.55	0.77*	0.80	115,098	149,152
X3y3	0.56	0.74*	0.77	141,261	190,366
X6y3	0.56	0.69	0.71	188,692	274,164
X9y3	0.57	0.65	0.67	213,209	329,160
X12y3	0.58	0.62	0.64	222,800	359,106
XHy3	0.59	0.56	0.60	228,862	411,511

* 表 AUC 或精確率達到 0.7

表 19：決策樹－貢獻分類各區間結果比較表。

區間代碼	AUC	一般客戶				高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數	精確率	召回率	實際購買人數	預測購買人數
X1y1	0.54	0.28	0.37	7,212	25,899	0.28	0.34	5,170	18,332
X2y1	0.54	0.20	0.31	10,377	51,461	0.28	0.31	10,981	39,360
X3y1	0.54	0.16	0.30	11,681	72,807	0.28	0.30	16,960	61,393
X6y1	0.55	0.11	0.24	9,065	81,995	0.29	0.30	25,546	88,702
X9y1	0.55	0.08	0.21	6,695	85,949	0.28	0.31	33,159	117,252
X12y1	0.56	0.05	0.18	4,015	77,150	0.30	0.30	37,667	126,120
X1y3	0.54	0.60	0.55	33,517	55,466	0.28	0.40	8,342	29,281
X2y3	0.56	0.58	0.56	59,676	103,138	0.31	0.41	15,637	49,904
X3y3	0.56	0.54	0.54	70,595	129,690	0.32	0.42	21,347	65,790
X6y3	0.57	0.48	0.48	88,255	182,389	0.34	0.40	33,312	99,136
X9y3	0.57	0.43	0.44	93,804	218,894	0.34	0.39	42,061	123,392
X12y3	0.57	0.39	0.41	91,963	235,713	0.35	0.38	47,718	135,742
XHy3	0.58	0.31	0.37	81,577	264,537	0.35	0.38	61,997	178,559

* 表 AUC 或精確率達到 0.7

表 20：決策樹－是否為高貢獻客戶各區間結果比較表。

區間代碼	AUC	高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.56	0.28	0.37	5,582	20,105
X2y1	0.56	0.27	0.33	11,738	42,926
X3y1	0.56	0.27	0.33	18,446	67,125
X6y1	0.56	0.28	0.32	27,229	95,603
X9y1	0.56	0.28	0.31	34,003	119,520
X12y1	0.57	0.30	0.32	39,874	134,882
XHy1	0.58	0.30	0.33	52,778	173,526
X1y3	0.58	0.29	0.42	8,749	30,597
X2y3	0.59	0.31	0.42	16,273	53,351
X3y3	0.60	0.32	0.43	22,342	70,921
X6y3	0.60	0.32	0.42	34,507	107,936
X9y3	0.60	0.33	0.41	44,722	136,032
X12y3	0.60	0.33	0.41	51,975	158,010
XHy3	0.60	0.32	0.43	71,328	221,638

* 表 AUC 或精確率達到 0.7

銷成本最低則可選擇 X2y3，若追求營收最大化則可採用 X6y3。

對貢獻分類的預測中，表 16 雖有多個區間的 AUC 大於 0.7，但精確率皆小於 0.7，一般客戶與高貢獻客戶甚至沒有同時高於 0.6 的情況，顯示無法準確預測貢獻分類。而表 17 對高貢獻客戶的預測能力較好，以 X12y3 精確率 0.69 為表現最佳。

4.3.2 決策樹

透過決策樹進行是否購買、貢獻分類和是否為高貢獻客戶的預測，結果記錄於表 18、於表 19 和表 20，運算時間依序為 6 分 39 秒、5 分 38 秒和 6 分 46 秒。然而在此實驗中，所有區間的 AUC 均小於 0.7，且預測貢獻分類和高貢獻客戶的精確率皆小於 0.5。因此，我們可以推斷這個決策樹可能不適合用來預測此問題。

4.3.3 XGBoost

以 XGBoost 進行預測所需的運算時間最長，對是否購買、貢獻分類和是否為高貢獻客戶分別用了 15 分 22 秒、33 分 23 秒和 15 分 9 秒，我們記錄結果於表 21、表 22 和表 23。針對預測是否購買，觀察於表 21 結果與邏輯斯迴歸於表 15 的結果相似，對於未來 1 個月是否購買的預測表現較差；預測未來 3 個月是否購買，以 X2y3 和 X3y3 這兩個區間的表現最佳，行銷規劃時如以成本為重可選擇 X2y3，營收為重則選擇 X3y3。而對於貢獻分類同樣無法準確預測，可從表 23 看出對於對高貢獻客戶的預測狀況較佳，AUC 皆大於 0.7，精確率以 X12y3 最高，但僅有 0.66。

4.3.4 LightGBM

最後使用 LightGBM 進行預測，對是否購買、貢獻分類和是否為高貢獻客戶的運算時間分別為 3 分 51 秒、3 分 53 秒和 3 分 47 秒，為速度最快的演算法。在預測是否購買方面，根據表 23 結果，針對未來 1 個月是否購買的預測表現，與上述演算法同樣表現不理想；預測未來 3 個月是否購買，則有四個區間屬於表現佳，包含 X2y3、X3y3、X6y3 和 X9y3，其 AUC 及精確率皆在 0.7 以上，行銷面以成本為重可選擇 X2y3，營收優先則選擇 X9y3。而貢獻分類的預測結果亦與上述演算法同樣表現不理想。表 28 看出對於對高貢獻客戶的預測狀況與 XGBoost 相似，AUC 皆大於 0.7，精確率以 X12y3 最高，但僅有 0.66。

表 21：XGBoost—是否購買各區間結果比較表。

區間代碼	AUC	未來會購買			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.65	0.51	0.79	27,472	53,451
X2y1	0.66	0.51	0.57	38,745	76,366
X3y1	0.67	0.49	0.52	49,547	101,622
X6y1	0.7*	0.51	0.42	52,458	103,826
X9y1	0.71*	0.52	0.37	52,115	100,976
X12y1	0.73*	0.52	0.36	53,317	103,402
XHy1	0.77*	0.52	0.32	50,952	98,612
X1y3	0.69	0.8*	0.99	80,784	101,145
X2y3	0.71*	0.77*	0.98	141,700	185,014
X3y3	0.72*	0.74*	0.97	177,654	241,565
X6y3	0.73*	0.69	0.93	248,884	358,409
X9y3	0.74*	0.68	0.83	267,205	390,760
X12y3	0.75*	0.69	0.76	266,167	385,129
XHy3	0.79*	0.68	0.72	276,520	407,935

* 表 AUC 或精確率達到 0.7

表 22：XGBoost—貢獻分類各區間結果比較表。

區間代碼	AUC	一般客戶				高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數	精確率	召回率	實際購買人數	預測購買人數
X1y1	0.66	0.35	0.33	6,488	18,774	0.44	0.35	5,365	12,302
X2y1	0.66	0.32	0.14	4,588	14,231	0.47	0.27	9,748	20,901
X3y1	0.67	0.28	0.12	4,755	17,219	0.47	0.26	14,782	31,323
X6y1	0.69	0.24	0.05	1,878	7,930	0.49	0.27	22,913	46,704
X9y1	0.71*	0.22	0.01	433	1,935	0.52	0.25	26,866	51,981
X12y1	0.72*	0.18	0.00	51	282	0.53	0.25	31,287	58,655
X1y3	0.69	0.63	0.90	54,871	87,213	0.52	0.34	7,028	13,604
X2y3	0.71*	0.60	0.88	92,810	155,328	0.52	0.38	14,642	28,007
X3y3	0.71*	0.57	0.84	110,579	194,696	0.51	0.42	21,569	42,101
X6y3	0.72*	0.53	0.74	137,097	259,518	0.54	0.42	34,372	63,995
X9y3	0.73*	0.51	0.58	123,421	243,719	0.56	0.40	43,425	77,356
X12y3	0.74*	0.49	0.50	110,520	224,106	0.58	0.39	49,547	85,653
XHy3	0.76*	0.45	0.37	81,819	181,946	0.55	0.46	75,280	136,849

* 表 AUC 或精確率達到 0.7

表 23：XGBoost—是否為高貢獻客戶各區間結果比較表。

區間代碼	AUC	高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.71*	0.55	0.18	2,797	5,054
X2y1	0.7*	0.57	0.15	5,409	9,557
X3y1	0.7*	0.56	0.15	8,634	15,452
X6y1	0.7*	0.58	0.16	14,015	24,290
X9y1	0.71*	0.60	0.16	16,938	28,354
X12y1	0.72*	0.61	0.16	19,952	32,721
XHy1	0.77*	0.60	0.20	32,242	53,812
X1y3	0.75*	0.55	0.30	6,217	11,335
X2y3	0.77*	0.56	0.33	12,702	22,613
X3y3	0.78*	0.57	0.35	17,854	31,101
X6y3	0.78*	0.62	0.33	26,897	43,694
X9y3	0.79*	0.65	0.30	32,277	49,559
X12y3	0.79*	0.66	0.29	37,236	56,013
XHy3	0.8*	0.63	0.34	56,324	88,899

* 表 AUC 或精確率達到 0.7

表 24：LightGBM—是否購買各區間結果比較表。

區間代碼	AUC	未來會購買			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.65	0.52	0.78	26,928	51,953
X2y1	0.66	0.51	0.56	38,315	75,310
X3y1	0.67	0.49	0.52	49,383	101,135
X6y1	0.7*	0.51	0.42	52,484	103,919
X9y1	0.71*	0.52	0.37	51,576	99,687
X12y1	0.73*	0.52	0.36	52,599	101,504
XHy1	0.77*	0.52	0.33	52,043	101,008
X1y3	0.69	0.8*	0.99	80,630	100,840
X2y3	0.71*	0.77*	0.98	141,494	184,549
X3y3	0.71*	0.74*	0.95	173,368	233,846
X6y3	0.71*	0.71*	0.87	233,182	330,558
X9y3	0.73*	0.7*	0.79	252,926	363,479
X12y3	0.75*	0.69	0.77	268,909	390,862
XHy3	0.79*	0.68	0.72	274,679	403,496

* 表 AUC 或精確率達到 0.7

表 25：LightGBM—貢獻分類各區間結果比較表。

區間代碼	AUC	一般客戶				高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數	精確率	召回率	實際購買人數	預測購買人數
X1y1	0.66	0.34	0.32	6,273	18,289	0.43	0.36	5,477	12,713
X2y1	0.66	0.32	0.15	4,794	15,193	0.46	0.28	10,026	21,718
X3y1	0.67	0.27	0.12	4,786	17,520	0.47	0.27	15,081	32,197
X6y1	0.69	0.24	0.06	2,243	9,468	0.49	0.27	23,094	47,115
X9y1	0.7*	0.21	0.02	681	3,279	0.52	0.25	26,484	51,019
X12y1	0.72*	0.19	0.01	203	1,046	0.54	0.25	30,968	57,712
X1y3	0.69	0.63	0.88	53,589	85,016	0.49	0.37	7,577	15,516
X2y3	0.7*	0.60	0.86	91,090	152,162	0.51	0.40	15,471	30,350
X3y3	0.71*	0.57	0.80	104,990	182,888	0.50	0.44	22,649	45,352
X6y3	0.71*	0.54	0.67	124,615	232,741	0.54	0.42	34,431	64,251
X9y3	0.72*	0.51	0.55	116,071	227,206	0.56	0.41	44,032	78,857
X12y3	0.74*	0.49	0.50	111,913	227,172	0.57	0.41	51,411	90,081
XHy3	0.76*	0.45	0.36	79,757	177,050	0.55	0.46	76,186	138,224

* 表 AUC 或精確率達到 0.7

表 26：LightGBM—是否為高貢獻客戶各區間結果比較表。

區間代碼	AUC	高貢獻客戶			
		精確率	召回率	實際購買人數	預測購買人數
X1y1	0.71*	0.54	0.20	3,020	5,622
X2y1	0.7*	0.56	0.16	5,701	10,189
X3y1	0.7*	0.55	0.16	9,115	16,500
X6y1	0.7*	0.57	0.17	14,335	25,048
X9y1	0.71*	0.60	0.16	16,891	28,337
X12y1	0.72*	0.61	0.16	20,094	33,044
XHy1	0.77*	0.60	0.20	32,648	54,445
X1y3	0.75*	0.52	0.32	6,685	12,873
X2y3	0.77*	0.55	0.35	13,436	24,623
X3y3	0.77*	0.57	0.35	18,162	31,959
X6y3	0.77*	0.63	0.32	26,124	41,775
X9y3	0.78*	0.65	0.31	33,306	51,437
X12y3	0.79*	0.66	0.31	38,912	59,040
XHy3	0.8*	0.63	0.35	57,650	91,290

* 表 AUC 或精確率達到 0.7

4.3.5 模型比較分析

本研究採用不同機器學習模型，對是否購買、貢獻分類及高貢獻客戶進行預測，比較 14 種區間的實驗結果。在預測是否購買的情境中，對於未來 1 個月皆無法準確預測；對於未來 3 個月行為，邏輯斯迴歸有 3 個最佳區間（X2y3、X3y3 和 X6y3），XGBoost 有 2 個最佳區間（X2y3 和 X3y3），而 LightGBM 則有 4 個（X2y3、X3y3、X6y3 和 X9y3）。對預測貢獻分類，所有演算法表現皆不盡理想，而預測高貢獻客戶以邏輯斯迴歸的 X12y3 最佳，雖其精確率僅達 0.69，但實務上仍有可用性。

總的來說，LightGBM 對未來 3 個月是否購買的預測表現最佳，其運算速度最快、可挑選區間最多，更能夠彈性因應市場變化。而邏輯斯迴歸預測是否購買的表現也不錯，亦有機會應用於預測客戶是否會成為高貢獻客戶。因此以下分別觀察兩演算法最佳區間的變數重要性。在變數重要性的判斷上，引用 LightGBM 內建的排序，然而邏輯斯迴歸的變數重要性，則是以變數標準化後所作的迴歸係數之絕對值來判斷，絕對值越大代表該變數對預測結果的影響越大。

首先觀察 LightGBM 模型預測是否購買，分別比較 X2y3、X3y3、X6y3 和 X9y3 區間。總變數共 32 個，變數重要性如表 27 所示，我們專注探討前 10 名（約 30%），其底色灰色者。前 3 個變數：歷史第 1 筆與特徵最後日間距天數、年齡、Recency 最近一次消費，為所有區間的前 4 名重要變數。總排序第 4 的變數為第 1 筆及最後 1 筆間距天數，此變數在 X2y3 排第 3 名、X3y3 排第 4 名、在 X6y3 和 X9y3 排第 6 名，因此對於前兩者較短資料區間而言，客戶持續關注品牌的時間相比較長資料區間的影響力更大。總排序第 5 為 Frequency 消費次數，此變數在 X2y3 排第 7 名，在其他區間皆為第 5 名，對於短資料區間影響力稍低，延長資料區間便有所上升。

總排序第 6 為平均折扣，此變數在 X2y3 第 8 名、X3y3 第 6 名、在 X6y3 和 X9y3 皆是第 4 名；另線上平均折扣總排序為第 7，在各區間排名隨時間增長而下降，門市平均折扣總排序 16，在 X9y3 排第 7，未排入其他區間的前 10 名。由此推論，對於較短資料區間而言，線上平均折扣相比平均折扣更重要，而對於較長資料區間來說，平均折扣則較為重要，特別是 X9y3 門市平均折扣的影響力高於線上平均折扣。

總排序第 8 為 Monetary 消費金額，此變數在 X2y3 第 10 名，在其他區間為第 7 或第 8 名；而線上和門市消費金額均未排入任一區間前 10 名。因此，總消費金額有其重要性，但區分通路後重要性下降。最後，有些變數僅在一些區間排入前 10 名，比如總排序第 9 的是女性品項購買數量，在前三個區間皆排入前 10，且對 X2y3 的影響力較大。總排序第 10 為廣宣頻繁活躍，此變數僅在前兩個資料區間排入第 9 名。總

表 27：LightGBM—是否購買－變數重要性排序。

總排序	特徵說明	X2y3	X3y3	X6y3	X9y3	排序總計
1	歷史第 1 筆與特徵最後日間距天數	1	1	1	2	5
2	年齡	2	2	2	3	9
3	Recency 最近一次消費	4	3	3	1	11
4	第 1 筆及最後 1 筆間距天數	3	4	6	6	19
5	Frequency 消費次數	7	5	5	5	22
6	平均折扣	8	6	4	4	22
7	線上 _ 平均折扣	5	7	9	9	30
8	Monetary 消費金額	10	8	7	8	33
9	購買數量 _ 商品 _ 女性	6	9	8	13	36
10	廣宣頻繁活躍	9	9	11	12	41
11	購買數量 _ 商品 _ 特殊品	12	10	13	10	45
12	線上 _M	13	13	12	11	49
13	線上 _ 第 1 筆及最後 1 筆間距天數	11	11	12	16	50
14	線上 _R	18	16	10	15	59
15	門市 _M	16	13	17	17	63
16	門市 _ 平均折扣	25	20	14	7	66
17	購買數量 _ 商品 _ 小孩	15	12	18	22	67
18	門市 _R	19	15	15	18	67
19	購買數量 _ 熱銷品	17	17	20	17	71
20	購買數量	20	18	14	20	72
21	購買數量 _ 週末	14	14	21	23	72
22	郵政頻率 _ 一般	31	19	16	14	80
23	購買數量 _ 商品 _ 運動	23	21	19	21	84
24	線上 _F	21	22	24	24	91
25	購買數量 _ 商品 _ 男性	27	24	22	19	92
26	門市 _ 第 1 筆及最後 1 筆間距天數	24	23	23	25	95
27	郵政頻率 _ 高頻	26	25	25	27	103
28	門市 _F	28	26	26	26	106
29	郵政頻率 _ 特殊	22	29	30	31	112
30	線上 _RFM 分群	31	26	27	28	112
31	門市 _RFM 分群	29	27	28	29	113
32	RFM 分群	30	28	29	30	117

排序第 11 的特殊品項購買數量，恰好在 X3y3 和 X9y3 排第 10 名。總排序第 14 的線上最近一次消費，則是僅在 X6y3 排第 10 名。

接著，觀察邏輯斯迴歸預測是否購買（X2y3、X3y3 和 X6y3）與預測高貢獻客戶（X12y3）的變數重要性，分別記錄於表 28 和表 29，其中灰底者為重要性前 10 名的變數。

表 28：邏輯斯迴歸－是否購買－變數重要性排序。

總排序	特徵說明	X2y3	X3y3	X6y3	排序總計
1	Frequency 消費次數	1	1	1	3
2	歷史第 1 筆與特徵最後日間距天數	2	2	2	6
3	線上 _ 第 1 筆及最後 1 筆間距天數	3	3	5	11
4	購買數量 _ 熱銷品	9	4	3	16
5	廣宣頻繁活躍	5	5	9	19
6	購買數量	8	7	8	23
7	購買數量 _ 商品 _ 女性	10	8	7	25
8	Recency 最近一次消費	14	9	4	27
9	門市 _ 第 1 筆及最後 1 筆間距天數	6	6	17	29
10	線上 _ 平均折扣	4	16	13	33
11	年齡	13	10	10	33
12	線上 _F	7	15	12	34
13	線上 _M	11	11	15	37
14	郵政頻率 _ 特殊	20	14	6	40
15	Monetary 消費金額	12	12	18	42
16	購買數量 _ 商品 _ 特殊品	18	13	11	42
17	購買數量 _ 商品 _ 小孩	21	17	14	52
18	平均折扣	17	18	23	58
19	門市 _F	15	23	20	58
20	線上 _R	24	24	16	64
21	購買數量 _ 商品 _ 運動	23	21	21	65
22	門市 _R	26	22	19	67
23	第 1 筆及最後 1 筆間距天數	19	26	28	73
24	門市 _ 平均折扣	22	20	31	73
25	購買數量 _ 商品 _ 男性	25	25	25	75
26	郵政頻率 _ 高頻	27	27	22	76
27	線上 _RFM 分群	16	32	29	77
28	門市 _RFM 分群	29	28	24	81
29	RFM 分群	32	19	32	83
30	門市 _M	28	29	27	84
31	郵政頻率 _ 一般	30	31	26	87
32	購買數量 _ 週末	31	30	30	91

表 28 前 2 個變數：Frequency 消費次數、歷史第 1 筆與特徵最後日間距天數，恰好為所有區間前 2 名的重要變數。總排序第 3 的變數為線上的第 1 筆及最後 1 筆間距天數，此變數皆為 X2y3 和 X3y3 的第 3 名，在 X6y3 排第 5 名。有趣的是，門市的第 1 筆及最後 1 筆間距天數為總排序第 9 名，在 X2y3 和 X3y3 排第 6 名，但未排入 X6y3 前 10 名；全通路的第 1 筆及最後 1 筆間距天數更是不屬全任一區間的前 10 名。

表 29：邏輯斯迴歸－是否為高貢獻客戶－變數重要性排序。

特徵說明	X12y3 排序
Frequency 消費次數	1
Monetary 消費金額	2
線上 _M	3
線上 _F	4
Recency 最近一次消費	5
門市 _F	6
線上 _ 平均折扣	7
門市 _M	8
購買數量 _ 熱銷品	9
郵政頻率 _ 特殊	10
線上 _ 第 1 筆及最後 1 筆間距天數	11
歷史第 1 筆與特徵最後日間距天數	12
第 1 筆及最後 1 筆間距天數	13
門市 _RFM 分群	14
線上 _R	15
購買數量 _ 商品 _ 小孩	16
年齡	17
購買數量 _ 商品 _ 特殊品	18
平均折扣	19
郵政頻率 _ 高頻	20
門市 _ 平均折扣	21
廣宣頻繁活躍	22
購買數量 _ 商品 _ 女性	23
RFM 分群	24
郵政頻率 _ 一般	25
門市 _ 第 1 筆及最後 1 筆間距天數	26
線上 _RFM 分群	27
購買數量 _ 商品 _ 男性	28
購買數量 _ 週末	29
購買數量	30
門市 _R	31
購買數量 _ 商品 _ 運動	32

因此對於前兩者較短資料區間而言，客戶分別於線上通路、門市通路關注品牌的時間，影響力都相比長資料區間更重要。

總排序第 4 為熱銷品購買數量，此變數在 X2y3 排第 9 名、X3y3 第 4 名、X6y3 第 3 名，其影響力隨資料區間拉長而增加。總排序第 5 為廣宣頻繁活躍，此變數為 X2y3 和 X3y3 排第 5 名、在 X6y3 則為第 9 名，因此當前對廣宣有活躍反應的客戶，

在短期未來較可能回購。總排序第 6 和第 7 分別為總購買數量及女性商品購買數量，於各區間排名為第 7 到第 10 之間，而女性商品購買數量此變數影響力隨資料區間拉長而稍有提升。

最後，有些變數僅在一些區間排入前 10 名，比如總排序第 8 為 Recency 最近一次消費，未列入 X2y3 前 10、於 X3y3 排第 9 名、X6y3 排第 4 名，也是影響力會隨著資料區間增加的變數。總排序第 10 為線上的平均折扣，僅在 X2y3 排名第 4 名，說明線上折扣對於短期行銷相當重要。總排序第 11 為年齡，恰好在 X3y3 和 X6y3 排第 10 名。總排序第 12 的線上消費次數，僅在 X2y3 排第 7 名。而總排序第 14 的特殊郵政頻率，則特別在 X6y3 排名第 6，看來對較長區間來說有其重要性。

表 29 為邏輯斯迴歸預測預測高貢獻客戶（X12y3）的變數重要性，部分前 10 名變數與是否購買重疊，最不同的是 Monetary 消費金額，此為重要變數排名第 2，區分通路後線上排第 3 名、門市排第 8 名。由此可見，客戶未來 3 個月是否能成為消費貢獻前 20% 的客戶，與過往 12 個月的消費貢獻息息相關，而線上的重要性又高於門市通路。

以上結果凸顯了 Covid-19 疫情使得消費者對線上購物需求大增，這在我們的變數重要性中得到驗證。正如第一章研究背景所述，即便是在疫情趨緩之後，電子商務銷售額仍持續成長。因此，企業應該注重線上通路的客戶經營，妥善運用數據分析客戶喜好，提供更加個人化的產品和服務，以強化全通路體驗，順應消費者偏好的演進。

5. 總結

在本研究中，我們對 H&M 資料進行了深入分析，探討了在不同時間區間內，影響消費者購買決策的關鍵因素，及這些因素如何塑造有效的行銷策略。此外，我們考量了制訂行銷計畫時，準確預測與成本效益間的平衡重要性，可根據不同的營運目標，如成本最小化或營收最大化的策略，進一步選擇要採用資料區間。

5.1 客戶行為分析

本研究透過比對兩時間區間資料（X3y3 及 X9y3），深入分析類別型和連續型特徵變數，不僅驗證了長短資料區間具有結構差異，也找到以下影響客戶行為的關鍵因素。

從客戶基本面來看，當客戶的郵政編碼屬於特殊頻率時，會減少未來購買的可能性。而屬於對廣宣頻繁活躍的客戶，有較高的機會於未來再購，且在短區間時更加明顯。至於年齡，在購買行為上雖於無顯著差異，但隨著資料區間拉長，仍會改變年齡

與客戶行為的分佈結構。

從交易數據面來看，購買數量也是分辨客戶行為的重點，舉例來說，總購買數量、熱銷品購買數量、週末購買數量、女性商品及特殊商品購買數量都是影響力相當高的因素。此外，客戶與品牌互動的情形也是重點，客戶接觸品牌的時間越長（越早期開始購買），或是在計算期間持續關注品牌，未來都越可能持續消費。

最後，關於 RFM 指標與分群，研究結果顯示最近購買（R）的客戶更有可能未來再次購買，甚至成為高貢獻客戶；並且消費次數（F）或消費金額（M）越高的客戶，未來回購機會也較高。就分群結果編號 0 到 3 來說，分群編號 0 皆為 R 大、F 及 M 小，分群編號 3 為 R 小、F 及 M 大，依序代表為許久未購買且貢獻度最低、稍久未購買且貢獻度低、前陣子有購買且貢獻度中、近期購買且貢獻度高。RFM 分群編號 3 的客戶更有機會回購，尤其是在較長資料區間時表現更加明顯。另對於較長資料區間來說，客戶過往在線上及門市的消費行為，對於未來表現的影響力高；然而對較短資料區間而言，客戶過往在門市的行為幾乎不具影響力。

5.2 購買行為預測

研究比較不同機器學習模型對於是否購買、貢獻分類及高貢獻客戶的預測能力。研究結果推薦應用的模型包含：以 LightGBM 預測未來 3 個月是否購買，資料區間可採用 X2y3、X3y3、X6y3 和 X9y3；以邏輯斯迴歸預測是否為高貢獻客戶，資料區間採用 X12y3。針對 LightGBM 的資料區間選擇，可從營運思維切入，若期望行銷成本最低選擇 X2y3，若追求營收最大化則可採用 X9y3。或是以目標營收反推預計行銷人數，若人數介於某兩個區間之間時，可選擇人數較多的區間，並從中挑選預測購買機率較高的客戶群體，以提高行銷活動效率。

模型不僅是預測客戶的行為結果，還有助於我們發現在敘述性統計中未能突顯出來的重要變數。例如在 LightGBM 模型預測未來 3 個月購買行為時，客戶年齡和平均折扣在多數資料區間的模型中皆為重要變數，尤其是平均折扣，我們發現在短期資料區間內，線上平均折扣的重要性大於整體平均折扣；而在較長的資料區間則相反。又如使用邏輯斯迴歸模型預測高貢獻客戶時，我們發現，區分線上或門市通路的變數，線上通路普遍比門市通路對模型的預測貢獻更大。這一發現提示我們，在制定行銷策略時應更加關注線上客戶群體，以針對性地提升線上行銷活動的效果。

5.3 顧客經營與行銷策略建議

從市場趨勢和數據探索的結果來看，我們建議企業應該重視並改善其全通路的服務體驗，特別是在線上通路方面。有效地運用數據來理解客戶的偏好，並基於這些洞見來制定的差異化行銷策略，有助於更全面地滿足顧客的需求。根據本研究的發現，我們進一步提出以下具體的實施建議：

1. RFM 分群的應用：利用 RFM 分析識別不同的客戶群體，尤其關注那些最近有購買行為且貢獻度高的客戶。此外，可排除購買率較低的客戶，如郵政編碼屬於特殊頻率者。而針對短期促銷活動，可優先針對廣宣頻繁活躍的客戶，他們有較高的再購機會。

2. 運用模型優化行銷策略：根據 LightGBM 和邏輯斯迴歸模型的預測結果，企業可以更精確地識別具有高購買機率或高貢獻潛力的客戶群體，並搭配前述應用進行精準行銷。

3. 重視客戶與品牌的長期互動：客戶與品牌互動的歷史長短對於預測其未來購買行為至關重要。企業應關注那些長期以來積極與品牌互動的客戶，透過提供持續的關懷和價值，以增強這些客戶的忠誠度和長期價值。

4. 深化線上通路的客戶體驗：鑑於線上通路在客戶購買決策中日益重要，企業不僅需加強線上銷售和行銷策略，更應致力於打造一個無縫且引人入勝的數位購物體驗。這可能包含運用 AI 技術打造即時的個人產品推薦、優化網站和 APP 的使用者介面，以及創造互動和社群化的購物環境。此外，提供即時客服支援、靈活的支付選項和流暢的物流服務也能進一步提升顧客滿意度，從而轉化為更高的忠誠度和重複購買率。

本研究在運用客戶基本資訊與交易資料方面取得了一定進展，未來的研究可以透過更全面的數據整合，來豐富我們對消費者行為的了解。進一步細化預測目標和應用集成學習方法有助於提升模型的預測能力和應對市場快速變化的能力。這些努力將引導未來的行銷策略更加精準和有效，為企業在競爭激烈的市場中保持優勢提供關鍵支持。

6. 致謝

感謝主編、匿名副主編和審稿委員的寶貴建議，讓本文得以修訂，使其更適合讀者閱讀。

參考文獻

[1] Anitha, P. and Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), pages 1785-1792.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

[3] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. URL: https://doi.org/10.1145/2939672.2939785

[4] Chevalier, S. (2022). *Retail e-commerce sales worldwide from 2014 to 2026*. Statista. URL: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales

[5] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction* (Vol. 2, pages 1-758). Springer.

[6] Hughes, A. M. (1994). *Strategic database marketing: The masterplan for starting and managing a profitable, customer-based marketing program*. Probus Publishing Co.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017). URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[8] Kotler, P., Kartajaya, H., and Setiawan, I. (2021). *Marketing 5.0: Technology for Humanity*. Wiley.

[9] Kumar, V. and Reinartz, W. (2018). *Customer Relationship Management: Concept, Strategy, and Tools* (3rd ed.). Springer.

[10] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pages 53-65.

[11] Smith, P. (2023). *H&M Group - Statistics & Facts*. Statista. URL: https://www.statista.com/topics/3733/handm-group/#topicOverview

[12] 夏梅雪（2021）。團購型消費者購物型態集群分析與分類預測之研究（碩士論文）。國立高雄科技大學行銷與流通管理系。

[13] 郭瀚揚（2019）。資料探勘應用之研究：零售業的 RFM 分析架構（碩士論文）。國立臺灣師範大學全球經營與策略研究所。

[14] 程美蘭（2018）。使用 RFM 擴充模型對客戶分群及行銷策略探討: 以共同基金為例（碩士論文）。實踐大學資訊科技與管理學系碩士在職專班。

[15] 鄭子萱（2021）。顧客終身價值分析：行銷隨機模型與機器學習方法之實證研究（碩士論文）。國立臺灣大學商學研究所。

[16] 蕭維嘉（2020）。OMO 情境下消費者零售通路選擇與消費行為樣態分析 以品牌 A 為例（碩士論文）。國立臺灣大學商學研究所。

[Received July 2024; accepted December 2024.]

Data-driven omnichannel customer analysis for H&M: an RFM approach to predicting purchase behavior

Wei-Hsin Chang and Nan-Cheng Su[†]

Department of Statistics, National Taipei University, Taipei, Taiwan

ABSTRACT

Modern businesses must effectively utilize customer data for analysis and prediction to better understand customer, thereby maximizing customer management benefits and providing a comprehensive omnichannel experience. This study leverages competition data from H&M published on the Kaggle, using transaction data to calculate RFM indicators and applying K-means clustering analysis for further segmentation. It also examines the impact of varying lengths of transaction records on the predictive ability of the model. By employing statistical analysis and machine learning methods, the study forecasts customer purchasing behavior over the next one and three months, categorizing them into general and high-contribution customers. Based on the findings, it is recommended that businesses effectively utilize RFM segmentation to identify customers, particularly focusing on those with recent purchases and high contributions. Additionally, the study suggests optimizing marketing strategies using models such as LightGBM and logistic regression to predict customers with high probability of purchase or potential for significant contribution. Lastly, beyond data application, businesses should also prioritize long-term customer interactions and enhance the digital experience.

Key words and phrases: Decision Tree, K-means, LightGBM, Logistic Regression, Omni-channel, RFM, XGBoost.

JEL classification: C45, M31.

[†]Corresponding to: Nan-Cheng Su
E-mail: sunanchen@gmail.com