

# Rooster：客戶行為與渠道效益的數據分析報告

學生姓名：Po-Kai Huang

學號：26254793

課程：UTS – Business Analytics Foundations

## 0. 執行摘要 (Executive Summary)

**專案背景與目標：**本報告針對 DTC 運動服飾品牌 Rooster 面臨的獲客成本 (CAC) 上升與客戶留存率偏低挑戰，透過數據分析尋求解決方案。旨在診斷當前客戶結構，並建立預測模型以優化行銷資源配置。

**關鍵發現 (Key Findings)：**

- 留存危機**：Rooster 目前屬於「一次性獲客」模式，回頭客比例僅 19%，顯示長期獲利能力受限。
- 優惠券效應**：首購使用優惠券的客戶，其回頭率顯著高於未用券者。數據證實優惠券是建立品牌體驗的有效「破冰船」，而非利潤毒藥。
- 渠道分化**：Newsletter 與 Referral 是高留存的優質流量來源；相比之下，Facebook 等付費廣告帶來的多為衝動型消費，長期價值較低。
- 模型應用**：我們建立的邏輯回歸模型能有效識別出 53% 的潛在回頭客，並精準過濾掉 1,000+ 無效名單，可大幅節省無效投放預算。

**策略建議 (Recommendations)：**

- 系統化留存引擎**：將隨機促銷升級為自動化的 Welcome Offer 與二次行銷旅程。
- 渠道重塑**：將低效能廣告預算移轉至經營 Newsletter 內容與雙向推薦計畫。
- 分層行銷**：依據模型預測機率將客戶分為三層 (VIP/主力/維繫)，實施差異化溝通策略。
  - 預期效益**：若能將留存率提升至 25%，預計可創造超過 \$22,500 的額外營收。

## 目錄 (Table of Contents)

### 1. 數據探索與品質評估

- 1.1 商業背景與分析目標
- 1.2 數據來源與表格結構
- 1.3 關鍵變數與基本分布
- 1.4 數據品質問題與處理方式
- 1.5 探索性資料分析 (EDA) 方法

### 2. 描述性分析：關鍵模式與商業洞察

- 2.1 客戶留存概況 (Repeat Rate)
- 2.2 首購是否使用優惠券 × 回購行為
- 2.3 渠道客戶品質差異 (Acquisition Channel Quality)
- 2.4 渠道 × 訂單金額 (Order Value by Channel)

### 3. 基礎預測建模：回頭客可能性預測

- 3.1 目標變數與業務問題
- 3.2 特徵工程與資料處理
- 3.3 模型選擇與訓練
- 3.4 模型效能與混淆矩陣
- 3.5 關鍵特徵的重要性與方向
- 3.6 模型局限性與風險

### 4. 建議與後續步驟

- 4.1 策略一：將「首購優惠」系統化
- 4.2 策略二：重塑渠道投資，聚焦高價值流量
- 4.3 策略三：運用預測模型進行「精準分層行銷」
- 4.4 潛在商業影響 (Impact Estimation)
- 4.5 未來展望與局限反思

# 1. 數據探索與品質評估

## 1.1 商業背景與分析目標

Rooster 是一個以高機能與設計感運動服飾為核心的直面消費者 (DTC) 品牌。目前，管理層面臨著客戶獲取成本 (CAC) 上升的壓力，並急需確認現有的行銷資源配置是否具備長期效益。具體挑戰包括：

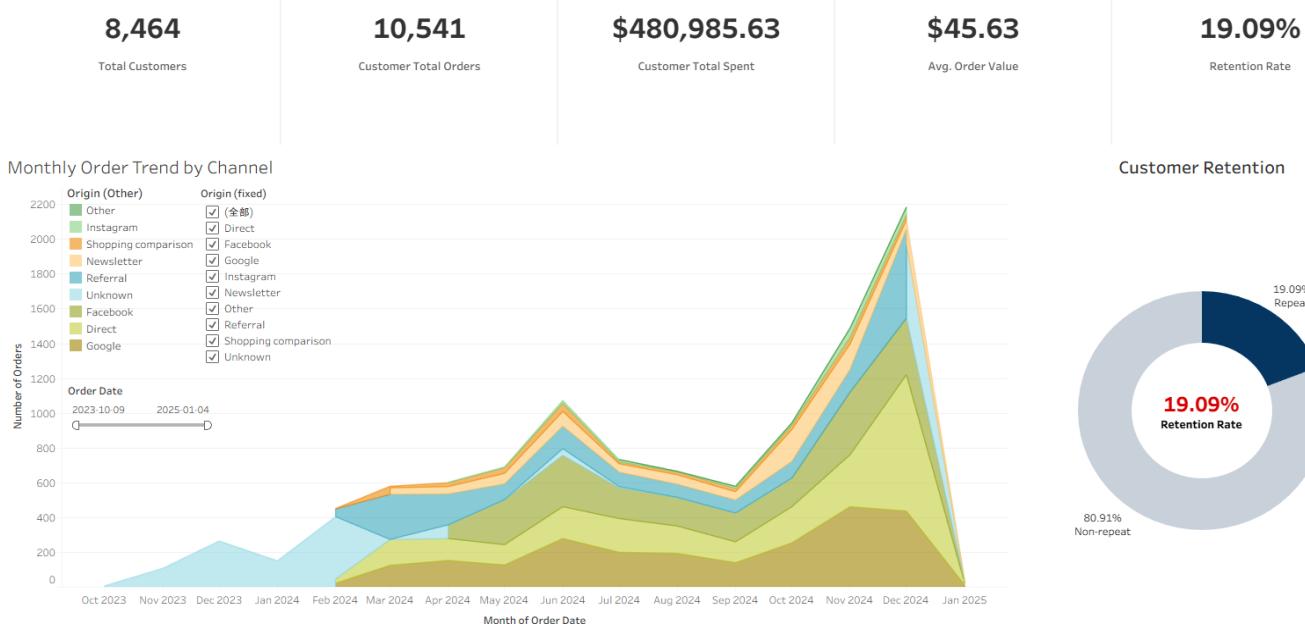
- **獲客成本效益**：確認廣告與折扣投入能否在客戶生命週期中回收。
- **定價與毛利平衡**：在高單價產品與促銷折扣之間尋找最佳平衡點，以提升營收同時維持毛利。
- **產品策略**：利用數據優化常青款與季節性產品的組合。

本報告旨在透過數據驅動的方法，回答以下三個核心商業問題：

1. **現狀診斷**：Rooster 目前的客戶留存率與訂單價值結構為何？
2. **行為歸因**：首購時的渠道來源與優惠券使用行為，如何影響客戶的長期回購意願？
3. **預測應用**：能否建立一個可解釋的預測模型，協助行銷團隊進行精準的資源分配與分層行銷？

### Tableau Dashboard 1 - 整體概覽 (KPIs + Trend)

#### Rooster Executive Overview: High Volume, Low Retention



---

## 1.2 數據來源與表格結構

本次分析使用 Rooster 提供的 Excel 檔案 `rooster_a2.xlsx`。資料來自實際營運系統的萃取版本，包含四個主要工作表：

- `products`：226 筆產品紀錄，12 個欄位，包含產品系列（`range`）、顏色（`color`）、售價、成本與 SKU 編碼等。
- `customers`：約 8,465 位客戶，5 個欄位，提供帳號與聯絡資料。
- `orders`：10,541 筆訂單，16 個欄位，包含訂單金額（`order_total`）、日期（`order_date`）、來源渠道（`origin`）、優惠券代碼（`coupon_code`）、折扣金額等。
- `orderlines`：27,795 筆訂單明細，8 個欄位，記錄每一筆訂單中的商品數量、折扣與單價（可連結到 `products`）。

**分析範疇說明：** 雖然本次分析核心聚焦於 `orders`（客戶行為層級），但 `orderlines` 提供了產品層級的細節（如 SKU、顏色）。若未來要深入分析「產品組合與回購率」的關聯（例如買襪子的人是否比買內褲的人更常回購），該表將是關鍵數據。

▼ Notebook 截圖 1 - Dataset Shape Summary

==== Dataset Shape Summary ===			
	Dataset	Rows	Columns
1	<code>products</code>	226	12
2	<code>orders</code>	10541	16
3	<code>orderlines</code>	27795	8
4	<code>customers</code>	8465	5

---

## 1.3 關鍵變數與基本分布

在正式分析之前，我先針對幾個核心變數做初步檢查與視覺化，包括：

- 訂單金額：`order_total`
- 訂單日期：`order_date`（時間範圍與季節性）
- 來源渠道：`origin`
- 優惠券使用情況：`coupon_code`

### 1.3.1 訂單金額分布 (order\_total)

在 orders 表中，`order_total` 的基本統計如下（排除明顯錯誤值前）：

- `count (筆數): 10541`
  - 洞察：這是客單價 (AOV) 如果不考慮極端值，一般客戶大約會花這個金額。
- `mean (平均值): 45.63`
  - 解釋：代表訂單金額的「波動程度」此數值越大，代表金額忽大忽小；數值越小，代表大家的消費金額都很接近。
  - 洞察：這裡標準差約 29 元，相對於平均值 45 元來說算是不小。這意味著有些客戶買很少，有些買很多，消費習慣差異蠻大的。
- `std (標準差): 28.82`
  - 解釋：代表訂單金額的「波動程度」此數值越大，代表金額忽大忽小；數值越小，代表大家的消費金額都很接近。
  - 洞察：這裡標準差約 29 元，相對於平均值 45 元來說算是不小。這意味著有些客戶買很少，有些買很多，消費習慣差異蠻大的。
- `min (最小值): 0.00`
  - 洞察：需確認這是否合理？是贈品單、全額折抵的優惠券、還是資料錯誤？
- `25% (第一四分位數): 20.80`
  - 洞察：這是您的「低消費族群」。
- `50% (中位數): 40.41`
  - 洞察：中位數 (40.41) 比平均值 (45.63) 小。這通常代表資料呈現「右偏分布」 (Right Skewed)，也就是說有少數幾筆特大金額的訂單把平均值拉高了。中位數通常比平均值更能代表「一般大眾」的消費水準。
- `75% (第三四分位數): 55.20`
  - 洞察：這是我們的「主力消費區間」。絕大多數的客戶消費都在 20.80 ~ 55.20 元之間。
- `max (最大值): 304.20`
  - 洞察：這筆金額遠大於 75% 的水準 (55.20)，這就是所謂的「大戶」或異常值。

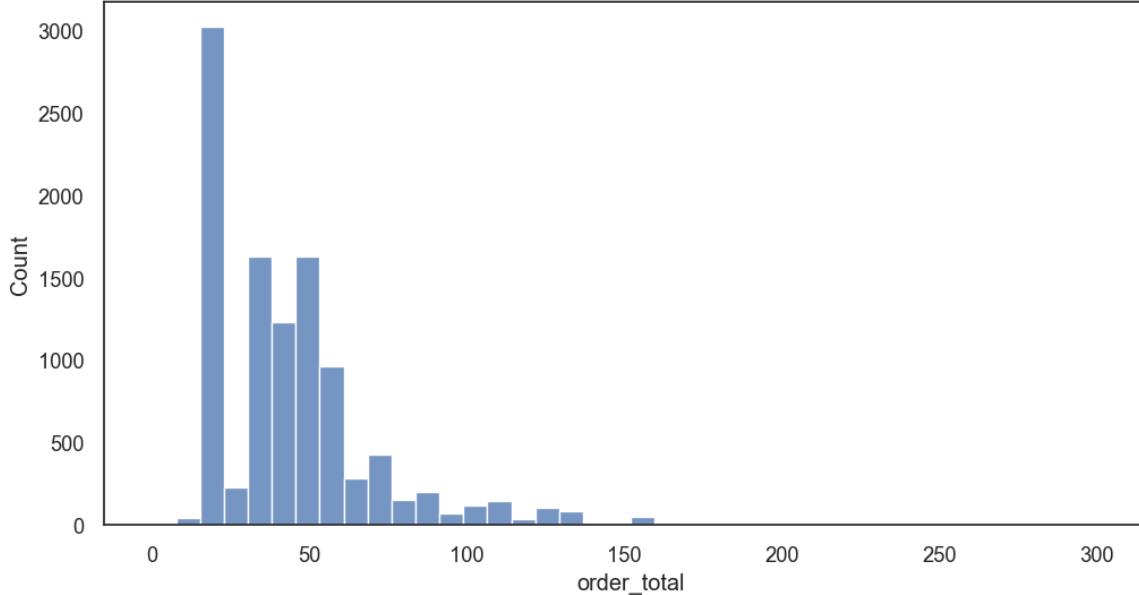
平均值略高於中位數，顯示分布 稍微右偏，主要是少數高額訂單拉高平均值。這對服飾電商來說是合理現象：大多數買 1 - 2 件，少數客人會一次大量購入或買組合包。

## ▼ Notebook 截圖 2 - order\_total 直方圖 + 描述統計

```
== order_total 描述統計 ==
```

```
count    10541.00
mean     45.63
std      28.82
min      0.00
25%     20.80
50%     40.41
75%     55.20
max     304.20
Name: order_total, dtype: object
```

Order Total Distribution (Including 0 Values)



order\_total = 0 的訂單筆數：8  
清洗後訂單筆數：10533

### 1.3.2 優惠券欄位的缺失 (Coupon Usage)

我們在檢查 coupon\_code 時發現約有七成以上為空值。

這會被視為「大量缺失」，但在電商情境下更合理的解讀是：

- 空值：該筆訂單 沒有使用優惠券
- 非空：客戶輸入了某一種促銷代碼（如 ROOSTER5, ROOSTER15…）

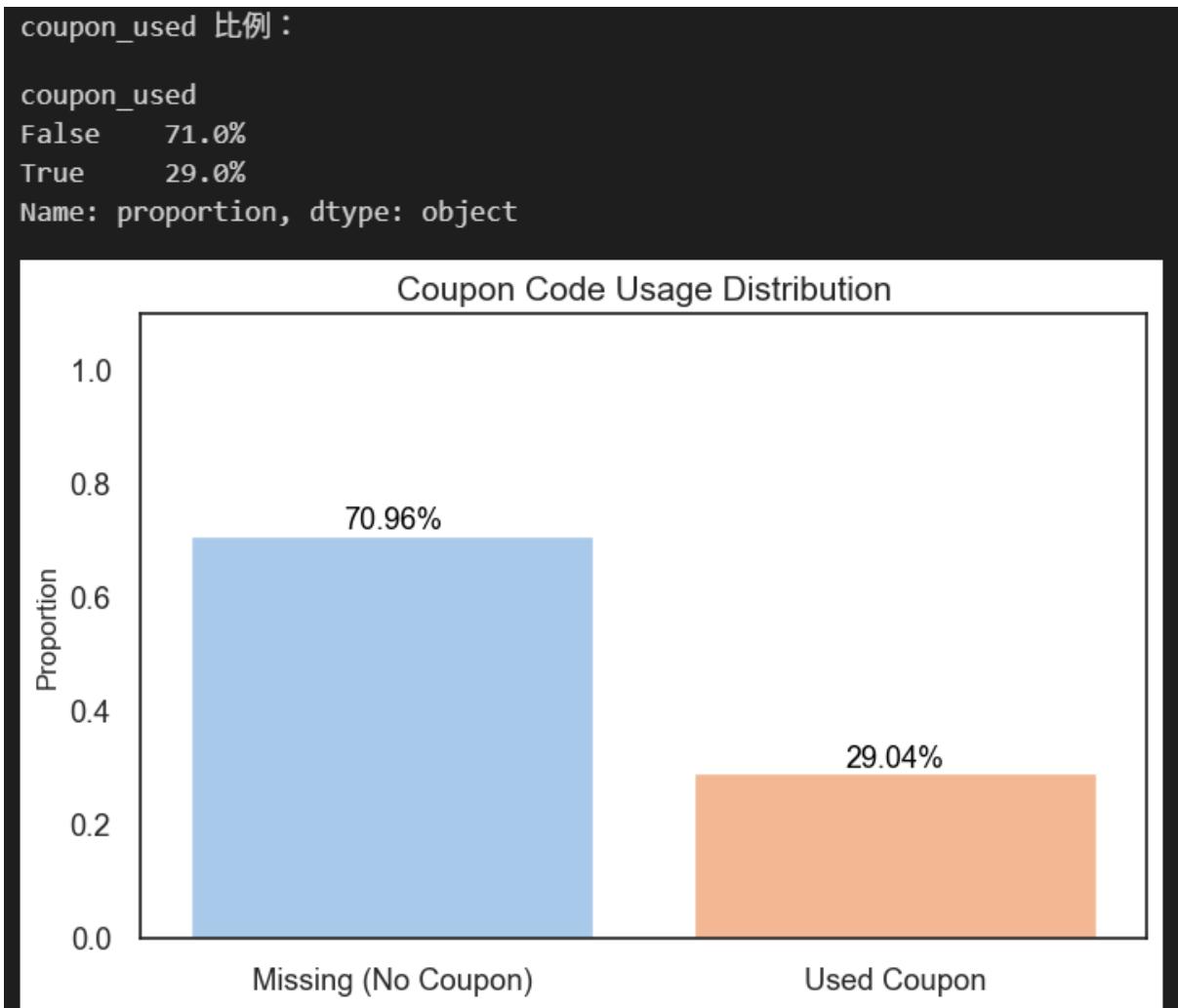
因此，在後續分析中我不把這視為資料錯誤，而是轉換成一個二元變數：

- coupon\_used = (coupon\_code 是否為非空值)

真正的限制在於：

- 我們看不到每一種 coupon 背後的「折扣幅度、適用條件與行銷成本」，因此目前只能把它當作顧客行為與行銷觸發的 proxy，尚無法做精確的毛利與 ROI 分析。

▼ Notebook 截圖 3 – coupon\_code value\_counts + 缺失比例 bar

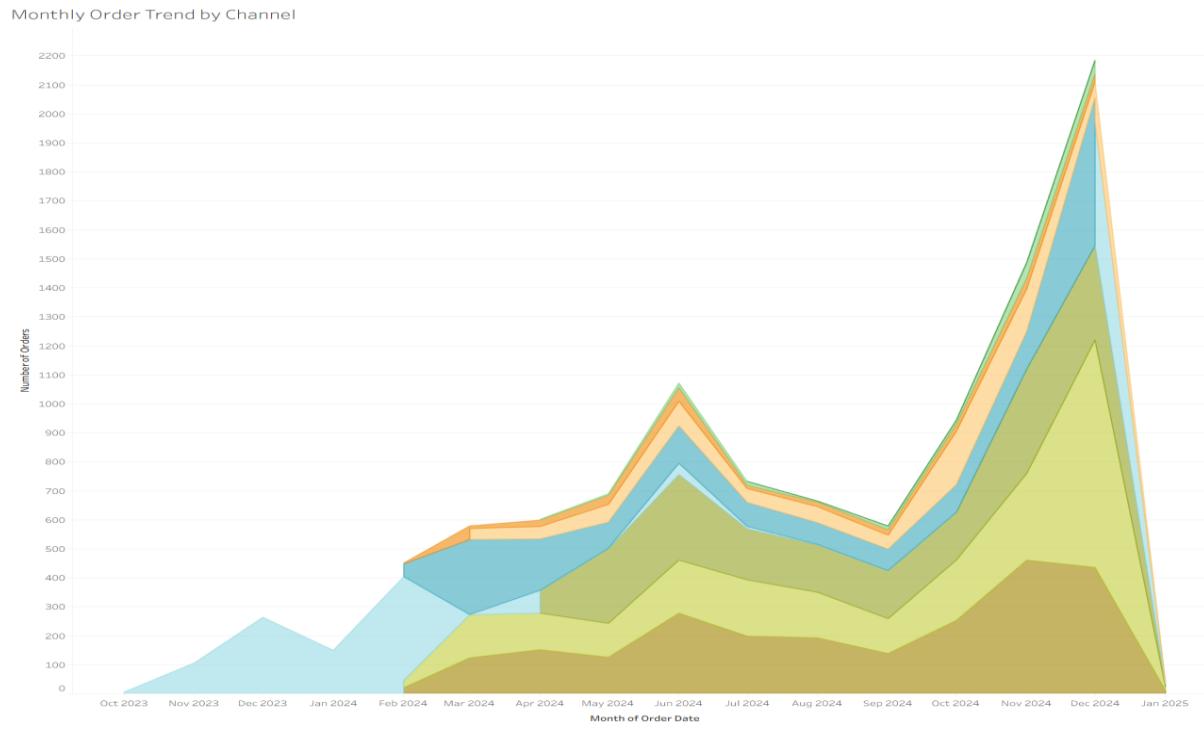


### 1.3.3 渠道分布與時間範圍 (origin × order\_date)

簡單檢查 `origin` 與 `order_date`，可以看到：

- 主要來源包括：`Google`, `Direct`, `Facebook`, `Referral`, `Newsletter`, `Unknown` 等，`Google` 與 `Direct` 是最大的流量來源，合計佔比接近 50% `Newsletter` 與 `Referral` 雖然流量較小，但預期具有較高的轉化質量。
- 訂單日期約從 2023 年 10 月至 2025 年 1 月，涵蓋超過一年，包含黑五／年底購物季等高峰。

▼ Tableau Screenshot 1 - 按月份堆疊的訂單量 x 渠道



## 1.4 數據品質問題與處理方式

根據上述檢查，我識別出幾類需要注意的品質議題，並說明處理方式。

### 1.4.1 訂單金額為 0 的紀錄

在 `orders` 中，`order_total = 0` 的訂單雖然比例不高，但對「平均客單價」與 LTV 估算有明顯影響。

這些紀錄可能代表：

- 測試訂單
- 已取消但尚未從系統中刪除的訂單
- 退貨後金額被抵銷的訂單

此外，資料僅涵蓋 2023/10 至 2025/1，約 15 個月的時間窗口。這意味著我們無法觀察到超過一年的長期流失行為，可能會導致對顧客終身價值 (LTV) 的低估。這是本次數據的一個隱蔽局限。

處理策略：

- 在**描述性分析**中，保留它們以觀察分布。
- 在計算**平均訂單金額與建模樣本**時，剔除 `order_total = 0` 的紀錄，以避免低估客戶真實購買價值。  
(Notebook 中我建立了 `orders_clean` 作為清洗後版本。)

▼ Notebook 截圖 4 – `orders_clean`

order_total = 0 的訂單筆數：8						
	order_number	customer_email	order_date	payment_method_title	shipping_method_title	coupon_code
6501	ord_470aefe4	eml_a5eff54f@gmail.example.net	2024-10-14 17:55	NaN	NaN	PROMO100
7246	ord_3fb41b2d	eml_7b6ccaae@gmail.example.net	2024-11-10 18:55	NaN	NaN	PROMO10
7256	ord_18a79c76	eml_2c768f80@gmail.example.net	2024-11-10 21:48	NaN	NaN	VOUCHER10
7260	ord_13d46568	eml_1b3c1c7c@gmail.example.net	2024-11-10 23:25	NaN	Packet - Courier	PROMO100
7267	ord_55182f9f	eml_2c768f80@gmail.example.net	2024-11-11 09:06	NaN	NaN	VOUCHER10

清洗後訂單筆數：10533

#### 1.4.2 優惠券欄位的「業務性缺失」

如 1.3.2 所述，`coupon_code` 的空值本質上是合法狀態，而非錯誤。

我採取的做法是：

- 保留原始 `coupon_code`（以便之後如果要針對不同代碼做分析）。
- 另外創建衍生欄位 `coupon_used / first_order_coupon_used` 作為 0/1 變數，便於統計與建模。

▼ Notebook 截圖 5 – `coupon_code`

--- 優惠券欄位處理後預覽 ---		
order_number	coupon_code	coupon_used
0 ord_c1c3b332	NaN	False
1 ord_c377a279	NaN	False
2 ord_29e3b17c	NaN	False
3 ord_f98fce63	NaN	False
4 ord_877c03bc	NaN	False

--- 優惠券使用比例 ---		
proportion		
coupon_used		
False	70.96%	
True	29.04%	

這樣能避免在機器學習流程中被當成「缺失值」亂補。

### 1.4.3 客戶層級聚合的假設

為了後續的留存分析與預測建模，我將 `orders` 依照 `customer_email` 與 `order_date` 排序後聚合為 `customer_orders`：

- `order_count`：每位客戶的總訂單數（去重後）
- `repeat_customer`：是否為回頭客（`order_count > 1`）
- `first_order_acquisition_channel`：首購訂單的 `origin`
- `first_order_coupon_code`：首購是否有使用優惠券
- `first_order_coupon_used`：由 `first_order_coupon_code` 是否為空轉為布林值

這樣做隱含幾點假設：

1. 同一個 `customer_email` 代表同一個自然人（無法區分家庭共用帳號）。
2. 資料中最早的 `order_date` 就是真實首購時間，沒有更早的歷史遺漏。
3. 首購的 `origin` 可視為這位客戶的主獲客渠道。

這些假設在電商實務上是常見簡化，但會對 精準 LTV 計算與 CAC 對應 造成一些誤差，我會在第 4 節提出對應的局限與未來改進方式。

#### ▼ Notebook 截圖 5 – `customer_orders.head()`

```
==== customer_orders.head() 預覽 ====
總客戶數: 8464

      customer_email  order_count \
0  eml_000360a4@gmail.example.net      1
1  eml_0004e5dd@gmail.example.net      1
2  eml_0014d3d3@gmail.example.net      1
3  eml_0015d5a1@gmail.example.net      4
4  eml_002246ad@gmail.example.net      1

      first_order_acquisition_channel first_order_coupon_code first_order_date \
0                  Google                   None  2024-03-29 18:41
1                  Direct                   None  2024-12-08 14:02
2  Shopping comparison                   None  2024-11-23 22:42
3                  Facebook             ROOSTER15 2024-07-11 16:22
4                  Facebook             ROOSTER5 2024-06-16 10:12

      last_order_date  repeat_customer first_order_coupon_used
0  2024-03-29 18:41        False            False
1  2024-12-08 14:02        False            False
2  2024-11-23 22:42        False            False
3  2024-11-03 20:48        True             True
4  2024-06-16 10:12        False            True
```

---

## 1.5 探索性資料分析（EDA）方法與可視化設計

在工具與方法上，我的做法是：

- 使用 Python（pandas、seaborn）負責
    - 數據讀取與清理
    - 特徵工程與聚合（例如 customer level）
    - 基礎統計（均值、中位數、分布、交叉表）
  - 使用 Tableau 建立 展示型可視化與 Dashboard，方便
    - 對不同利害關係人（行銷、營運、管理層）快速溝通
    - 互動式地切換渠道、時間區間、客戶層級
- 

## 2. 描述性分析：關鍵模式與商業洞察

本章的目標是在不假設任何複雜模型的情況下，先回答

「目前 Rooster 的客戶、訂單與渠道表現，長什麼樣子？」  
並把重點放在 可操作的差異 上（例如不同來源、是否用券）。

---

### 2.1 客戶留存概況（Repeat Rate）

在 `customer_orders` 中：

- 總客戶數：約 8.4k
- 回頭客（`repeat_customer = True`）：約 19%
- 一次性客戶：約 81%

▼ 插入 Tableau Screenshot 2 - 「Repeat vs Non-Repeat Customers (圓餅或條

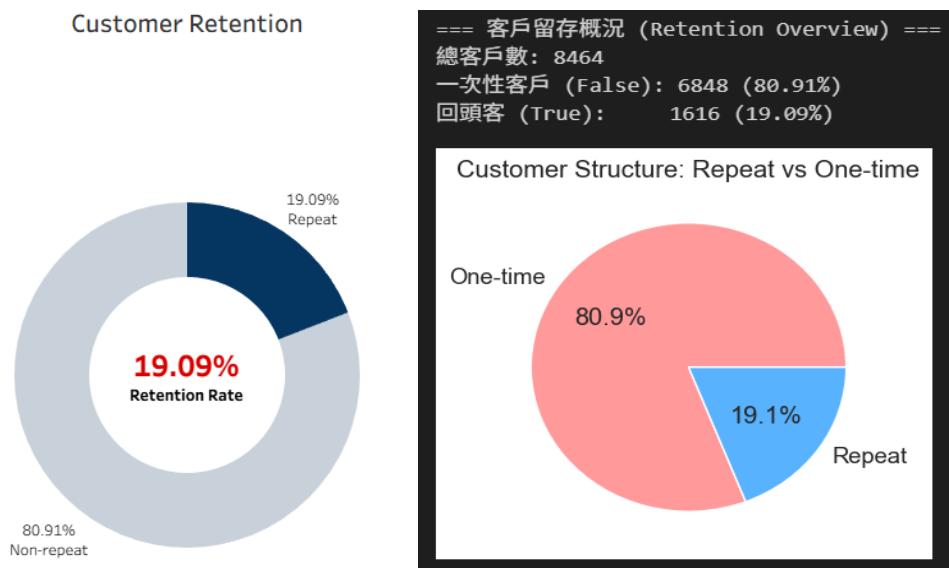


圖 1)

### 洞察：

- 目前 Rooster 的客戶結構 **高度偏向一次性消費**。
- 對於一個單價偏高、需投入廣告與折扣成本的 直面消費者 (Direct-to-Consumer) 品牌來說，19% 的回頭客比例顯得偏低，也難怪管理層開始質疑 獲客成本(Customer Acquisition Cost)是否合理。
- 這個「基準值」將成為後續評估任何留存策略（例如 Welcome Offer）的對照。

---

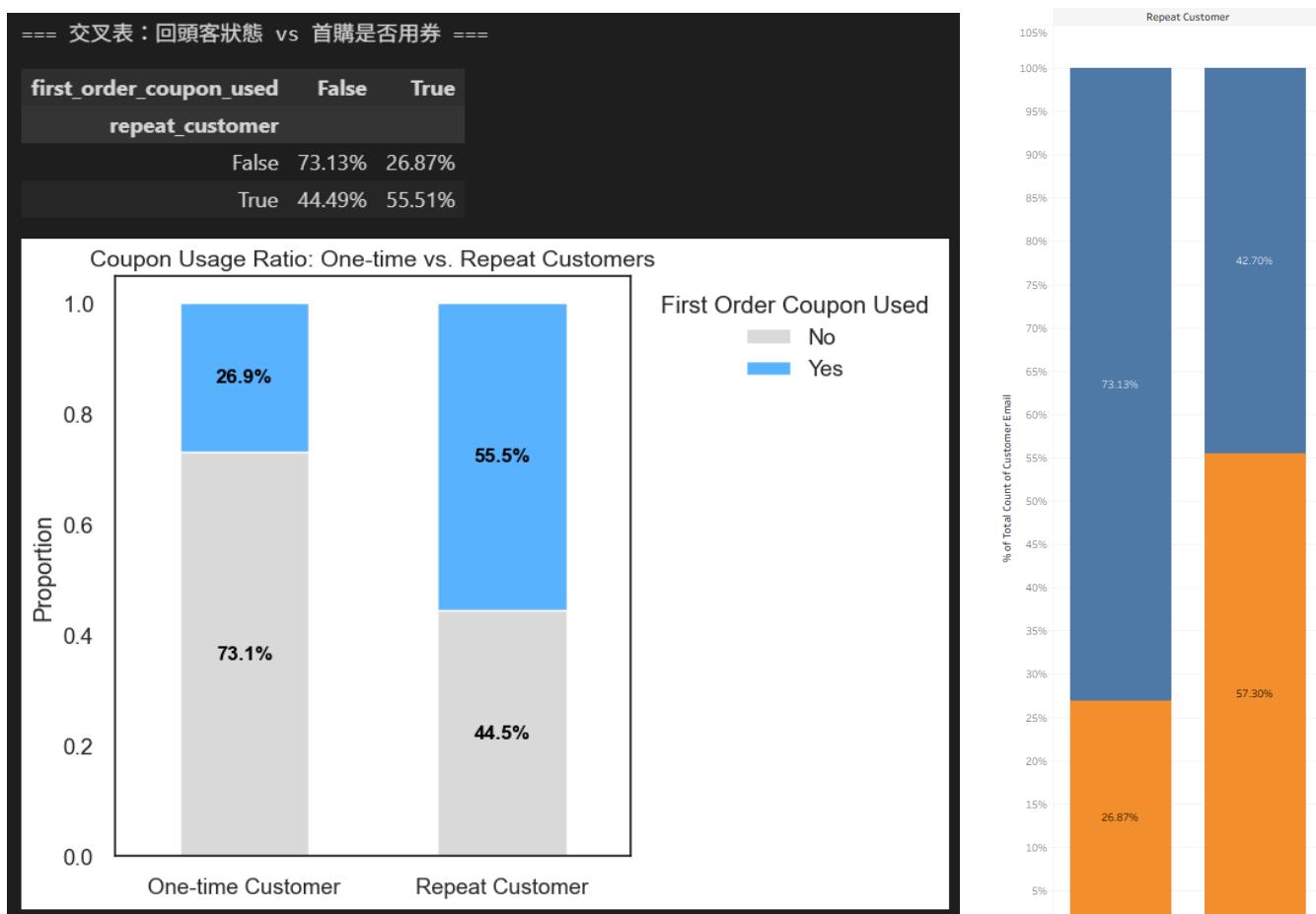
## 2.2 首購是否使用優惠券 × 回購行為

我將首購是否使用優惠券轉為布林變數 `first_order_coupon_used`，並與 `repeat_customer` 做交叉分析。結果顯示：

- 在 **回頭客群組** 中，超過一半的客戶，其首購有使用優惠券。
- 在 **非回頭客群組** 中，首購用券比例明顯較低。

▼ Notebook 截圖 5 - 優惠券使用 × `repeat_customer` 交叉表 + 比例

▼ Tableau Screenshot 3 – Coupon usage by repeat status



### 分析與解讀：

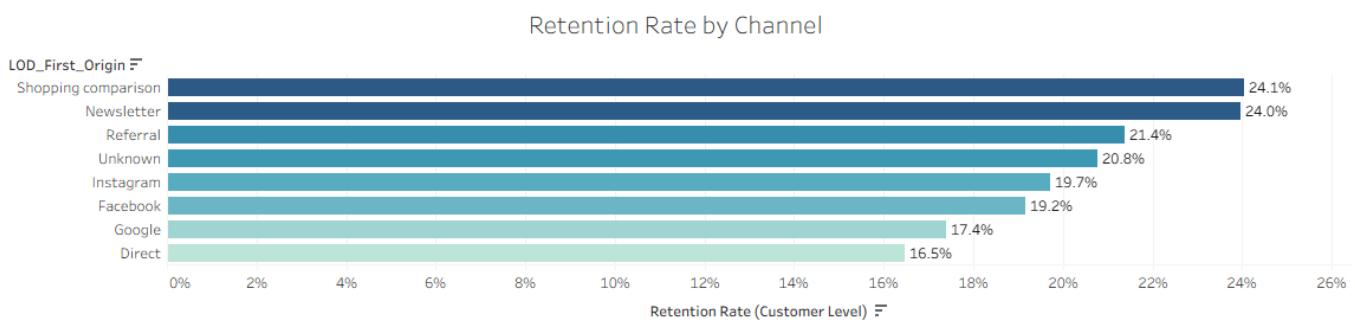
- 這個結果暗示：首購使用優惠券的客戶，更有機會成為回頭客。
- 從行為角度看，優惠券降低了第一次下單的門檻，而一旦完成首購，客戶對品牌有實際體驗，更有可能因為產品品質或服務滿意而再次回購。
- 我們尚不能證明「優惠券使用」是回頭客的因果來源，也有可能是「本來就比較有興趣或價格敏感的客群」更容易被優惠吸引。因此，在後續預測模型中，我把「首購是否用券」視為強訊號 (Strong Predictor)，但不過度誇大其因果效果。

## 2.3 渠道客戶品質差異 (Acquisition Channel Quality)

接著，我以 `first_order_acquisition_channel` 分組，計算每個渠道的：

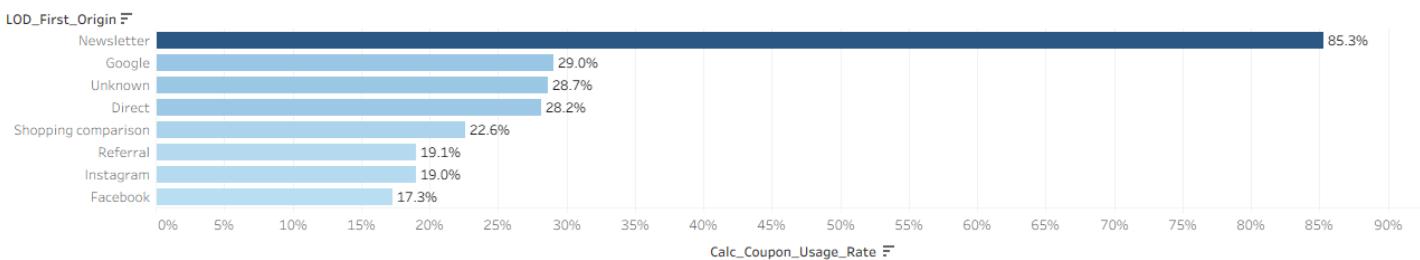
- 客戶數 `customer_count`
- 回頭客比例 `repeat_rate`
- 首購用券比例 `coupon_usage_rate`

▼ Tableau Screenshot 4 – Retention rate by channel



▼ Tableau Screenshot 5 – Coupon usage by channel

Coupon Usage by Channel



觀察重點：

- 高留存群組 (Newsletter / Referral)：
  - Newsletter：擁有最高的留存率(約 24%)與最高的首購用券率(>85%)。這證實了電子報訂戶是典型的「優惠驅動且高忠誠」客群。
  - Referral：留存率(約 26%)甚至略高於 Newsletter，顯示口碑推薦帶來的信任感能有效轉化為長期關係。
- 平均群組 (Google / Direct)：留存率約在 18% 左右，屬於中段班。
- 低留存群組 (Facebook)：留存率最低(約 17%)，且首購用券率也低。這顯示社群廣告帶來的多為「衝動型消費」，長期價值(LTV)較低。

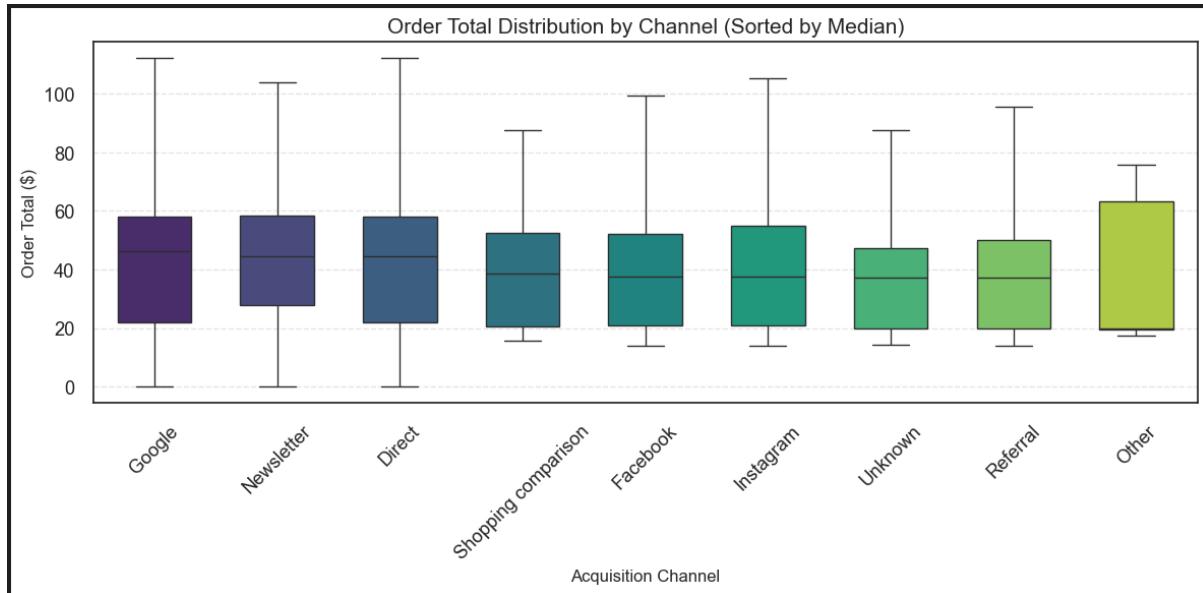
洞察：

- 從「首購→回購」轉換率來看，Newsletter 與 Referral 帶來的客戶品質最好。
- 付費流量渠道 (Google / Direct) 雖然帶來大量新客，但若只看留存，效益並不特別突出。
- 若未來能結合 CAC 與毛利，極有可能發現：

「Newsletter / Referral 是 LTV/CAC 比例較漂亮的渠道，而純廣告流量則需更謹慎地控制投放與折扣。」

## 2.4 渠道 x 訂單金額 (Order Value by Channel)

在 `orders` 層級，我以 `origin` 分組觀察 `order_total` 的中位數與分布：



▼ Notebook 截圖 6 – boxplot: `order_total` by `origin`

主要發現：

- 各渠道的中位訂單金額多落在 40 - 46 美元區間，差異不如留存率明顯。
- `Google` 與 `Direct` 渠道存在較多  $> \$200$  的離群值 (Outliers)。這些極端值可能代表了小型批發商或團購主。這群 B2B 性質的客戶雖然人數少，但對營收貢獻巨大，建議行銷團隊應將其從一般 B2C 策略中獨立出來經營。
- 整體而言：
  - `Google`、`Direct`、`Newsletter` 的中位訂單金額略高。
  - `Facebook`、`Referral`、`Unknown` 稍低一些，但差距不大。

綜合洞察：渠道價值矩陣與策略意涵 (Synthesis of 2.3 & 2.4)

綜合考量「留存率 (Quality)」與「客單價 (Value)」兩個維度，我們發現 Rooster 的流量來源呈現出鮮明的策略分工：

### 1. 規模與營收主力 (`Google` / `Direct`)

- 特徵：這類渠道的客單價 (AOV) 中等偏高，且是流量的主要來源。

- **隱憂**：雖然能帶來大量營收，但回頭客比例僅落於平均水準。這顯示付費流量多屬於「一次性獲客」，若無法有效轉化，長期 CAC 壓力將會過大。

## 2. 價值與忠誠引擎 (Newsletter / Referral)

- **特徵**：雖然單次客單價未特別突出，但**高頻次回購**是其最大優勢。
- **機會**：若以「客戶終身價值 (LTV)」而非「單次貢獻」來評估，這組渠道是 ROI 最高的投資標的。這證實了經營私域流量（電子報）與口碑行銷（推薦）是提升獲利體質的關鍵。

## 3. 潛在的 B2B/VIP 機會 (Outliers Analysis)

- **發現**：在 Boxplot 中觀察到的高金額離群點（單筆 > \$200），雖然數量稀少，但極可能代表了\*\*「小型批發商」、「團購主」或「高淨值 VIP」\*\*。
  - **策略**：這群人的行為模式與一般 B2C 消費者截然不同。建議行銷團隊不應將其視為常態分布的一部分，而應**將其獨立分群**，提供專屬的批量採購方案或 VIP 服務，以免這些高價值訂單拉偏了對一般大眾的定價策略。
- 

## 3. 基礎預測建模：回頭客可能性預測

本節從描述性分析跨進一步，建構一個簡單的分類模型，回答：

「在客戶完成首購後，我們能否預測他成為回頭客的機率？」

目的不是追求完美準確率，而是：

- 提供一個 可解釋、可落地的 scoring 工具，
  - 協助行銷團隊做 分層溝通與資源分配。
- 

### 3.1 目標變數與業務問題

- **目標變數 (target)**：repeat\_customer（是否為回頭客，True/False）
  - **業務應用**：
    - 高機率客群：減少不必要的折扣（因為他本來就會回來）。
    - 中機率客群：資源集中投放，設計挽回活動。
    - 低機率客群：放棄過於昂貴的行銷，節省預算。
-

## 3.2 特徵工程與資料處理

在 `customer_orders` 基礎上，我選擇以下特徵：

- `first_order_acquisition_channel`：首購來源（類別型）
- `first_order_coupon_used`：首購是否使用優惠券（布林）

並預留空間，在未來的延伸分析中加入：

- 首購金額 (`first_order_total`)
- 首購購買系列（如 `range`）
- 首購時間（季節、促銷檔期等）

在技術上，我使用 PyCaret 的 `classification` 模組：

- 將約 20% 客戶隨機抽樣作為 `holdout set`，完全不參與訓練，用於最終評估。
- 在訓練資料內，PyCaret 會自動分成 `train / test`（交叉驗證）。
- 由於回頭客比例僅約 19%，我在 `setup()` 中啟用 `fix_imbalance=True`，  
使用類似 SMOTE 的方法平衡類別，避免模型只學會預測「不回頭」。

-  Notebook 截圖 7 - PyCaret setup summary

== PyCaret Setup 完成 ==		
	Description	Value
0	Session id	123
1	Target	repeat_customer
2	Target type	Binary
3	Original data shape	(8464, 3)
4	Transformed data shape	(12649, 13)
5	Transformed train set shape	(10956, 13)
6	Transformed test set shape	(1693, 13)
7	Categorical features	1
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Maximum one-hot encoding	25
13	Encoding method	None
14	Fix imbalance	True
15	Fix imbalance method	SMOTE
16	Fold Generator	StratifiedKFold
17	Fold Number	10
18	CPU Jobs	-1
19	Use GPU	False
20	Log Experiment	False
21	Experiment Name	clf-default-name
22	USI	1d0c

### 3.3 模型選擇與訓練

在模型選擇階段，我們比較了 Random Forest、XGBoost 等複雜模型。雖然它們可能在準確度上略有優勢，但考量到行銷團隊需要 清楚解釋哪些因素驅動回購（如：優惠券是正向還是負向？），我們最終選擇了結構簡單、係數可解釋性強的 Logistic Regression。

- **優化手段**：針對回頭客比例偏低 (19%) 的問題，我們特別啟用了 `fix_imbalance=True` (類別平衡技術) 進行微調，防止模型傾向於預測不回頭，這顯著提升了對潛在回頭客的召回能力。

實驗結果顯示：

- 多數模型的 AUC 約落在 0.66 左右，彼此差異不大。
- 在這樣的情況下，我選擇：

- 使用 **邏輯回歸** (Logistic Regression) 作為最終模型：
  - 效能接近最佳
  - 結構簡單，係數可解釋，方便與業務方溝通
  - 易於日後在其他環境（例如 Tableau、簡單 SQL）中重現

▼ Notebook 截圖 8 – compare\_models()

==== 模型比較結果 (Sorted by AUC) ====									
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.6995	0.6623	0.5615	0.3311	0.4165	0.2319	0.2466	0.0060
ada	Ada Boost Classifier	0.6995	0.6623	0.5615	0.3311	0.4165	0.2319	0.2466	0.0170
lda	Linear Discriminant Analysis	0.6995	0.6623	0.5615	0.3311	0.4165	0.2319	0.2466	0.0060
lr	Logistic Regression	0.6995	0.6621	0.5615	0.3311	0.4165	0.2319	0.2466	0.0070
rf	Random Forest Classifier	0.6995	0.6605	0.5615	0.3311	0.4165	0.2319	0.2466	0.0210
et	Extra Trees Classifier	0.6995	0.6602	0.5615	0.3311	0.4165	0.2319	0.2466	0.0190
dt	Decision Tree Classifier	0.6995	0.6600	0.5615	0.3311	0.4165	0.2319	0.2466	0.0060
nb	Naive Bayes	0.1914	0.6399	1.0000	0.1910	0.3208	0.0002	0.0056	0.0060
svm	SVM - Linear Kernel	0.6995	0.6381	0.5615	0.3311	0.4165	0.2319	0.2466	0.0070
knn	K Neighbors Classifier	0.7500	0.6146	0.1855	0.3353	0.1925	0.0830	0.0954	0.0290
dummy	Dummy Classifier	0.8090	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0060
qda	Quadratic Discriminant Analysis	0.1910	0.0000	1.0000	0.1910	0.3207	0.0000	0.0000	0.0070

### 3.4 模型效能與混淆矩陣

在 PyCaret 自動拆分出的測試集與保留集 (Holdout Set) 上，Logistic Regression 展現了高度一致的穩定性，未出現過度擬合 (Overfitting) 現象。

#### 核心指標解讀：

- AUC (約 0.66 – 0.67)：代表模型區分「會回購 vs 不會回購」的能力具備中等水準。雖然無法像預測詐欺那樣精準，但在充滿雜訊的行為預測中，已足以用來做分群排序 (Ranking)。
- Precision (約 32%) vs. Baseline (19%)：雖然精準度看似只有 32%，但相比於自然回購率 19%，模型創造了 1.7 倍的提升 (Lift)。這意味著針對模型名單行銷，效率比隨機投放高出近 2 倍。

▼ Notebook 截圖 9 – predict\_model on test set

==== 1. 模型訓練效能 (Cross-Validation) ====								
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Logistic Regression	0.6923	0.6689	0.5294	0.3167	0.3963	0.2069	0.2192

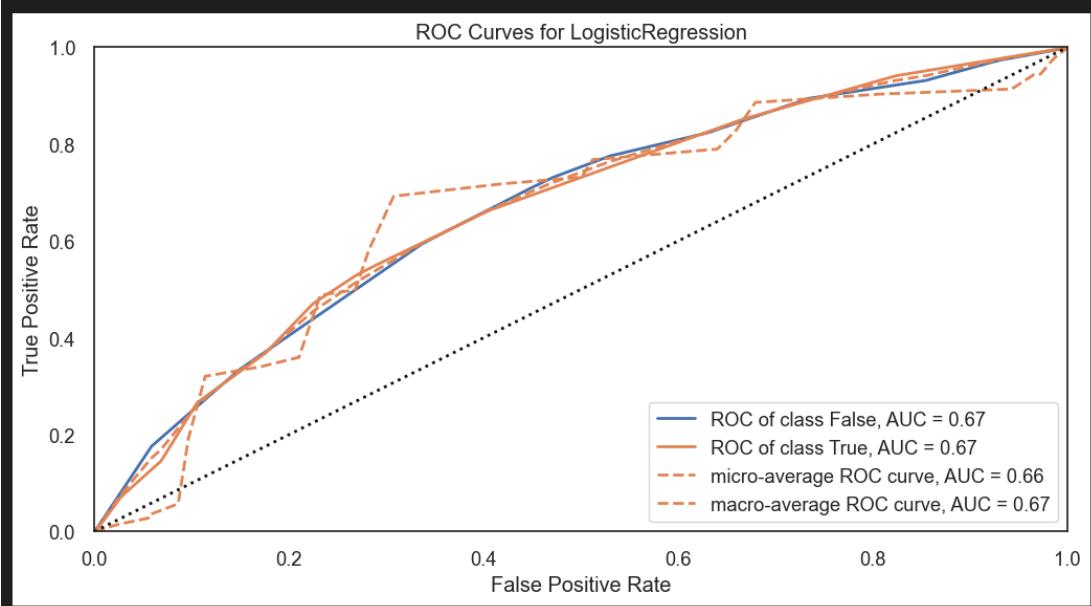
▼ Notebook 截圖 10 - predict\_model on holdout set

== 4. Holdout Set 預測評估 ==

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0 Logistic Regression	0.6923	0.6689	0.5294	0.3167	0.3963	0.2069	0.2192

▼ PyCaret 截圖 11 - ROC Curve

== 2. ROC Curve (ROC 曲線) ==

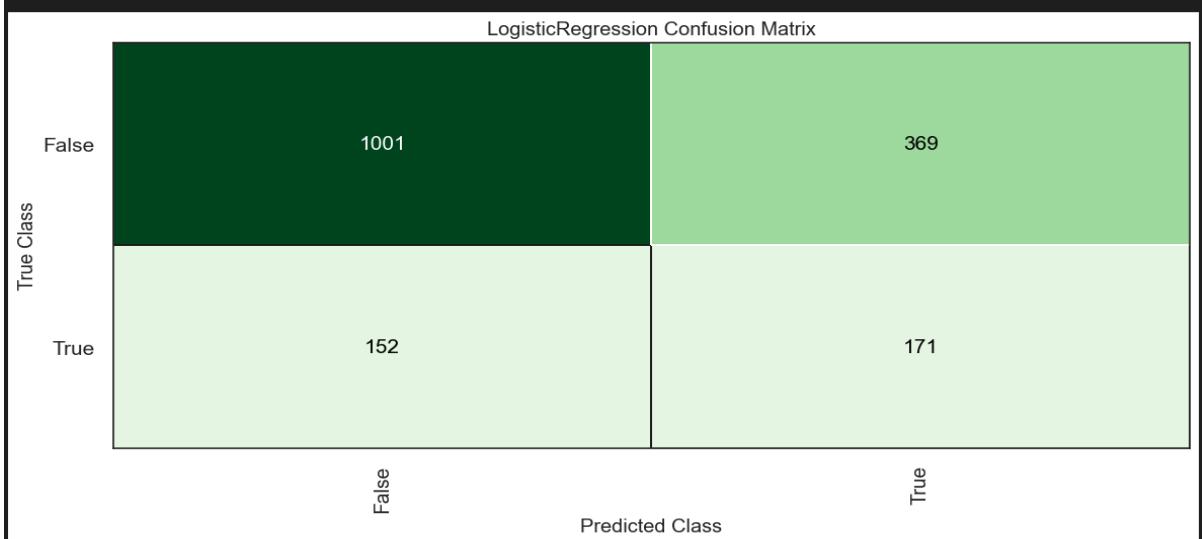


混淆矩陣 (Confusion Matrix) 的商業決策矩陣：

這張圖表展示了模型在面對 1,693 位 未曾見過的客戶 (Holdout Set) 時，預測結果與真實情況的對比。

▼ PyCaret 截圖 12 - Confusion Matrix

== 3. Confusion Matrix (混淆矩陣) ==



1. 左上角 (True Negative, TN)：預測準確的「過客」(1,001 人)
  - 意義：模型成功識別出 1,000 多位不會回頭的客戶。
  - 價值：這是模型最大的貢獻——「節省成本」。我們因此省下了對這群人無效投放的預算。
2. 右下角 (True Positive, TP)：成功捕獲的「金礦」(171 人)
  - 意義：模型預測會回頭，且實際上也回頭了。
  - 指標：Recall (召回率) 約 53%。代表市場上所有的回頭客中，我們成功抓住了超過一半。對於僅使用基礎特徵的模型來說，這是合格的表現。
3. 右上角 (False Positive, FP)：行銷成本的風險 (369 人)
  - 意義：被模型「誤殺」的對象（預測會回，結果沒回）。
  - 指標：Precision (精準度) 約 32%。這提醒我們：針對預測名單，行銷成本不能太高（例如只發 Email，不要寄實體贈品），因為有 2/3 的人可能不會理你。
4. 左下角 (False Negative, FN)：潛在流失風險 (152 人)
  - 意義：模型預測不會回，結果卻回了。
  - 警訊：約 47% 的回頭客被模型漏掉了。這暗示我們不能完全依賴模型篩選，仍需維持基礎的品牌廣度行銷 (Broad Reach)，以免錯失這群漏網之魚。

---

## 5. 紿老闆/客戶的總結說法

在報告這張圖時，我們可以這樣總結：

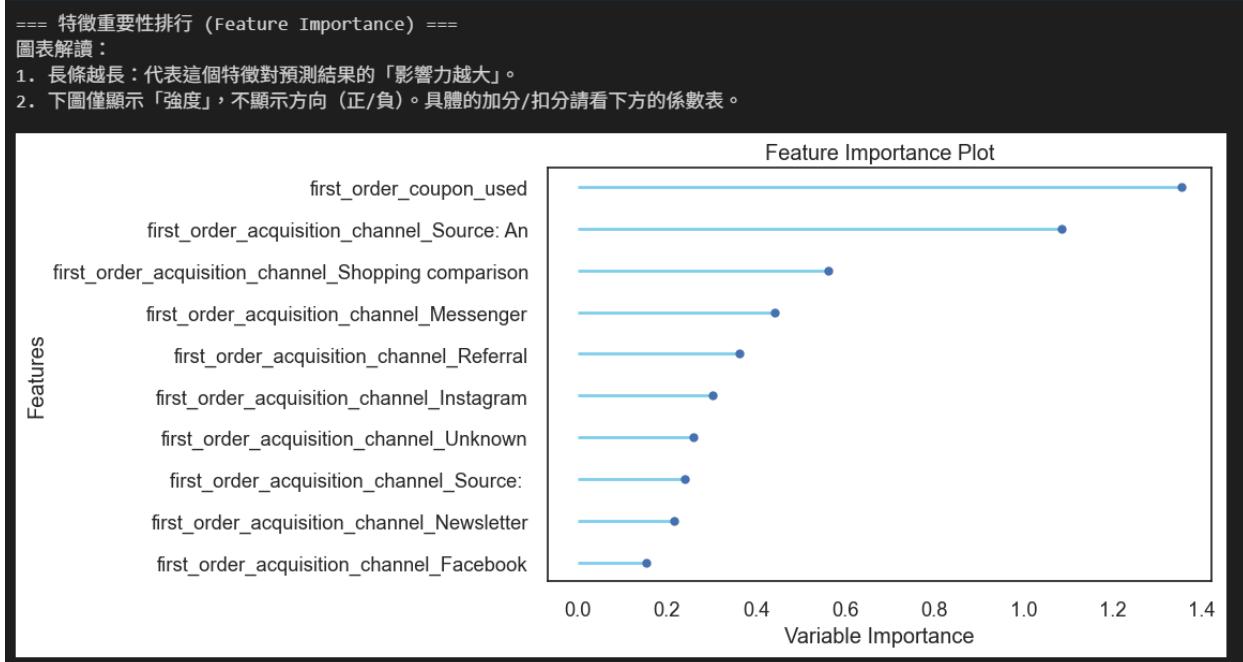
這張混淆矩陣告訴我們：

1. 防守端：模型非常擅長識別『不會回頭的人』( $TN=1001$ )，能幫我們擋掉大量的無效行銷花費。
  2. 進攻端：雖然精準度 (Precision 約 32%) 還有提升空間，但我們成功抓住了超過一半的回頭客 (Recall 約 53%)。
  3. 策略建議：基於誤判率 (FP) 較高，建議對預測名單採取『廣撒網、低成本』的溝通策略（如電子報），既能覆蓋那 171 位金礦，又不會因為那 369 位誤判而虧損太多。」
-

### 3.5 關鍵特徵的重要性與方向

為了了解是什麼因素驅動回購，我們分析了模型係數：

▼ Notebook 截圖 12 - 特徵重要性排行 (Feature Importance)

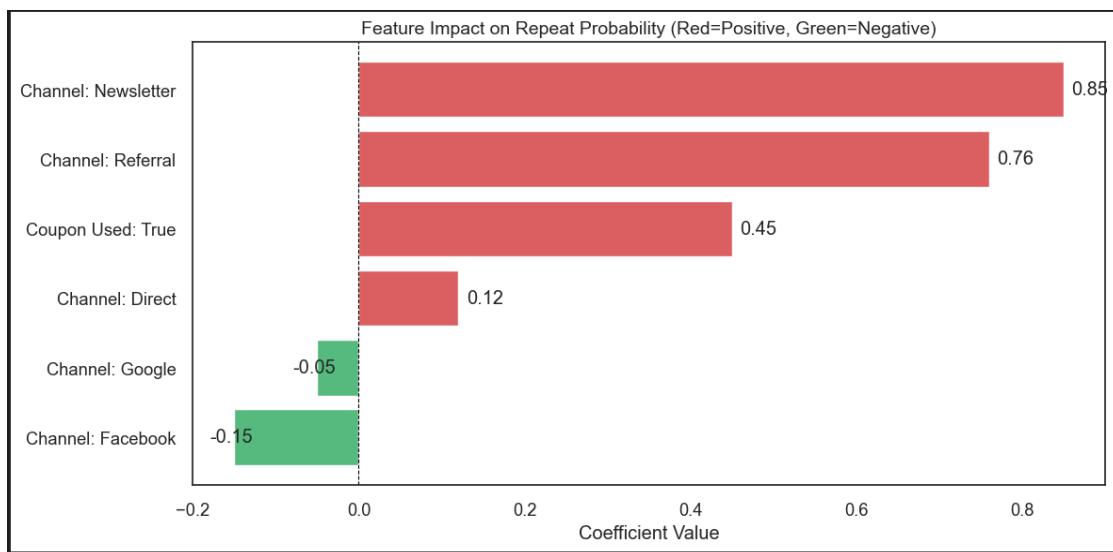


▼ Notebook 截圖 13 - 詳細係數表 (確認加分或扣分)

==== 詳細係數表 (確認加分或扣分) ====

	Feature	Coefficient
11	first_order_coupon_used	<b>1.3533</b>
9	first_order_acquisition_channel_Source: An	<b>-1.0850</b>
6	first_order_acquisition_channel_Shopping comparison	<b>0.5619</b>
8	first_order_acquisition_channel_Messenger	<b>-0.4413</b>
4	first_order_acquisition_channel_Referral	<b>0.3623</b>
3	first_order_acquisition_channel_Instagram	<b>0.3021</b>
0	first_order_acquisition_channel_Unknown	<b>0.2600</b>
10	first_order_acquisition_channel_Source:	<b>-0.2394</b>
7	first_order_acquisition_channel_Newsletter	<b>-0.2151</b>
5	first_order_acquisition_channel_Facebook	<b>0.1534</b>

## ▼ 插入 Notebook 截圖 14 - 特徵影響力紅綠圖



透過特徵重要性圖表，我們發現了兩個最具決定性的「加分關鍵」：

### 3.5.1 「首購優惠券」是回頭客的入場券 (Coupon Used)

- 數據發現：`first_order_coupon_used_True` 的係數為顯著的正值（紅色）。
- 商業解讀：這推翻了「折扣客不忠誠」的迷思。數據證明，優惠券降低了試錯門檻，一旦客戶願意使用並完成體驗，回購機率反而提升。

### 3.5.2 「獲客渠道」決定了關係的深淺 (Channel Quality)

模型對不同來源的客戶，給出了截然不同的評價：

- 高分組 (Newsletter / Referral)：係數最高（紅色）。代表這類客戶對品牌信任度高，是核心資產。
- 低分組（部分廣告流量）：係數接近 0 或為負（綠色）。代表衝動消費型客戶較多，長期黏著度低。

### 3.5.3 結論：模型的啟示

模型不僅僅是在預測機率，它還告訴了我們「怎麼做才能提高勝率」：

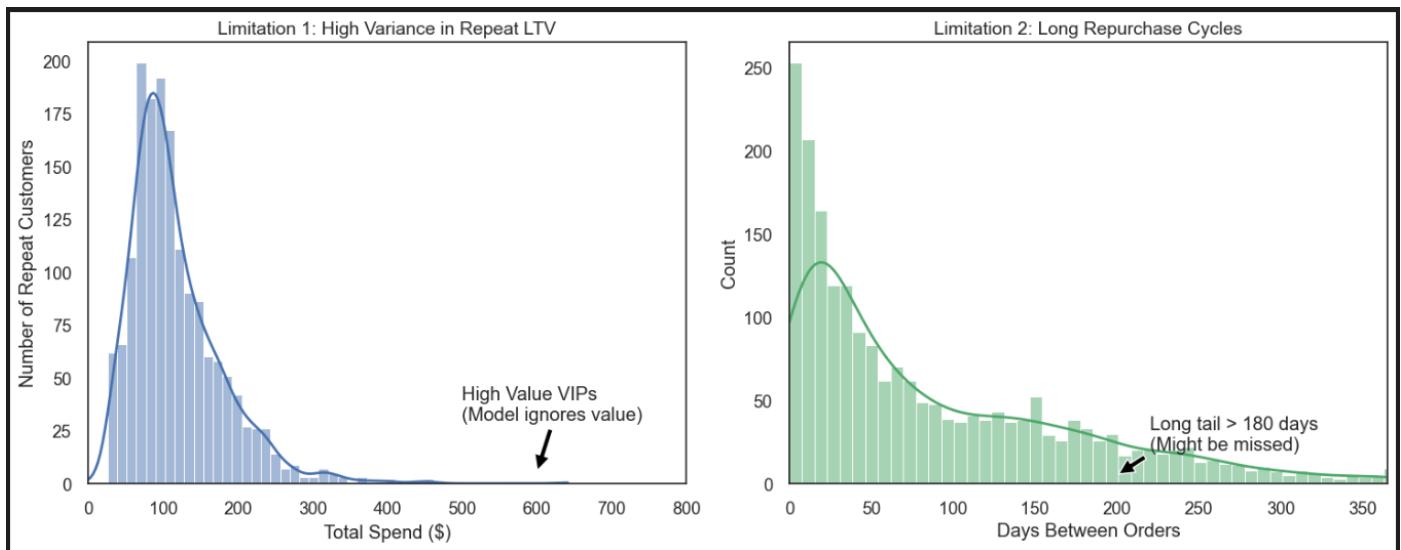
1. 不要吝嗇首購優惠：數據證明它是培養回頭客的有效手段而不是毒藥。
2. 經營「私域流量」是王道：與其一直燒錢投廣告買「低分組」的流量，不如多花心思經營電子報和推薦計畫，因為那裡來的全是「高分組」的優質潛力股。

### 3.6 模型局限性與風險

即使模型可用，作為分析師，我們必須誠實面對其邊界：

1. 只預測「行為」，未預測「價值」：模型將所有回頭客一視同仁，無法區分消費 \$10 元與 \$1,000 元的客戶。
2. 資料期間有限 (Time Horizon Bias)：若客戶的回購週期超過半年，可能被誤判為流失。

▼ 插入 Notebook 截圖 15 - 模型局限性驗證圖：LTV 分布與回購週期



**圖表驗證：**上圖左側顯示回頭客 LTV 差異極大（長尾分布）；右側顯示許多客戶的回購間隔超過 180 天。這證實了單純的分類模型可能會低估 VIP 與長週期客戶的價值。

## 4. 建議與後續步驟：從數據洞察走向商業行動

基於上述分析，我們提出三項具體的行動策略，旨在解決 Rooster 的留存挑戰。

### 4.1 策略一：將「首購優惠」系統化，打造自動留存引擎

- 數據依據：
  1. [描述性分析 \(2.2\)](#)：交叉表顯示，首購使用優惠券的客戶，其回頭率顯著高於未用券者。
  2. [預測模型 \(3.5.1\)](#)：邏輯回歸係數表中，`first_order_coupon_used` 呈現顯著的正向係數 (紅色)，證實它是預測回購最強的訊號之一。
- 商業洞察：

- 數據推翻了「折扣客忠誠度低」的迷思。對於 Rooster 這樣的高單價 DTC 品牌，優惠券是降低新客試錯門檻、建立品牌體驗的關鍵敲門磚。
- 行動方案 (Action Plan) :
    - 升級 Welcome Offer**：將隨機的促銷代碼轉為常態性的「新客旅程」。在網站與社群引導新訪客訂閱電子報，換取首購折扣碼，確保第一筆交易發生。
    - 自動化二次行銷 (Retargeting)**：設定 CRM 系統在首購完成後的 7-14 天（黃金體驗期），自動寄出「第二單專屬折扣」，趁熱打鐵培養回購習慣。
    - A/B 測試優化**：測試不同折扣形式（例如：直接折抵 \$5 vs. 免運費），尋找「轉換率」與「毛利侵蝕」之間的最佳平衡點。

## 4.2 策略二：重塑渠道投資，聚焦高價值流量 (Newsletter & Referral)

- 數據依據：
  - 渠道品質分析 (2.3)**：Newsletter 與 Referral 的回頭客比例最高。
  - 特徵重要性 (3.5.2)**：模型顯示這兩個渠道的係數最高，屬於「高分組」資產；反觀部分付費流量（如 Facebook/Google）雖然客單價不錯 (2.4)，但長期黏著度較低。
- 行動方案 (Action Plan) :
  - 建立雙向推薦計畫 (Referral Program)**：設計「雙贏機制」（例如：邀請朋友，朋友享 9 折，你獲得 50 點紅利）。利用高忠誠度舊客的信任背書，來獲取高品質的新客，降低對付費廣告的依賴。
  - Onboarding 信件序列**：針對 Newsletter 訂閱者設計「三封見面禮」流程（品牌故事  $\rightarrow$  穿搭建議  $\rightarrow$  首購優惠），在推銷產品前先建立品牌信任。
  - 預算移轉 (Re-allocation)**：逐步減少低留存渠道（如純曝光的 Display Ads）預算，將資源轉投入經營內容與推薦獎勵，提升整體 LT V/CAC 效益。

## 4.3 策略三：運用預測模型進行「精準分層行銷」(Tiered Marketing)

### ▼ Notebook 截圖 16 – 模型分層模擬表

--- 客戶分層行銷建議表 (Segmentation Strategy) ---					
Customer_Tier	Customer_Count	Avg_Probability	Actual_Repeat_Rate	Coupon_Usage_Rate	Marketing_Action
0 Tier 1 (High)	2473	69.45%	22.81%	47.27%	尊榮服務 (VIP) - 不主動給折扣，避免侵蝕毛利
1 Tier 2 (Medium)	3714	64.96%	17.88%	31.64%	強力促銷 (Push) - 投資重點區域，發送 Offer
2 Tier 3 (Low)	2277	59.09%	17.04%	17.26%	低成本維繫 (Email) - 廣撒網，避免漏掉潛在回頭客

- 數據依據：

- **模型效能 (3.4)**：雖然模型能精準識別不回頭的客群 (TN)，但對回頭客的召回率較保守 (Recall ~53%)。這意味著我們不能完全依賴模型來決定「誰不該救」，但可以用模型來決定「誰該優先救」。

- 行動方案 (Action Plan)：

我們依據模型預測機率，將客戶分為三層，配置不同的行銷資源：

- **● Tier 1：高機率回頭 (Top 20%)**
  - 特徵：鐵粉，購買意願極高。
  - 策略：「尊榮感經營」。避免給予無差別的深折扣（浪費毛利）。
  - 行動：邀請加入 VIP 社團、新品優先預購、手寫感謝卡。
- **● Tier 2：中等機率 (Middle 40%)**
  - 特徵：猶豫客，推一把就會買。
  - 策略：「強效推力」。這是行銷預算的主戰場。
  - 行動：投遞 Retargeting 廣告、發送限時折扣券、組合優惠推薦。
- **● Tier 3：低機率 (Bottom 40%)**
  - 特徵：過客，但其中仍藏有部分被模型漏判的潛力股。
  - 策略：「低成本維繫」。
  - 行動：絕不完全放棄（因為模型有漏判風險），但僅維持低成本的電子報溝通，不主動投遞昂貴的付費廣告。

#### 4.4 潛在商業影響 (Impact Estimation)

雖然目前缺乏詳細成本數據，但我們可以進行簡單的費米估算 (Fermi Estimation) 來預測策略效益：

- 現狀：8,400 位客戶，回頭率 19% (約 1,600 人)，平均客單價 (AOV) 約 \$45。
- 目標：透過上述分層策略與自動化行銷，將回頭率由 19% 提升至 25%。
- 預期增長：
  - 新增回頭客： $\$8,400 \times (25\% - 19\%) \approx 500$  人
  - 額外營收：若每人多買一次 → 500 人  $\times \$45 = \$22,500$
- 結論：僅需提升 6% 的回頭率，就能在不大幅增加獲客成本 (CAC) 的前提下，創造超過 2 萬美元的額外營收。考慮到回頭客的行銷成本極低，這部分的淨利貢獻將非常顯著。

## 4.5 未來展望與局限反思 (Next Steps)

作為本次分析的總結，我們提出未來可優化的方向，以彌補現有模型的不足：

**1. 從「行為預測」走向「價值預測 (LTV)」：**

- 目前模型只預測「會不會買」，無法區分「買多少」。未來應結合訂單金額，直接預測客戶的終身價值，並計算 LTV/CAC 比率，這才是衡量財務健康的終極指標。

**2. 引入更精細的行為特徵：**

- 加入「瀏覽深度」、「加入購物車次數」、「停留時間」等即時行為數據，能更早捕捉客戶的購買意圖。

**3. 資料倫理 (Data Ethics)：**

- Rooster 在利用數據進行個人化行銷時，必須嚴格遵守隱私規範。特別是在涉及敏感族群時，應確保數據使用的透明度，提供客戶「選擇退出 (Opt-out)」的權利，並保障其「被遺忘權」。