關鍵報告

(BDA Final Project)

Group 12

徐嬿鎔 (B04702077)

鄭皓 (B04702012)

廖文豪 (B04702010)

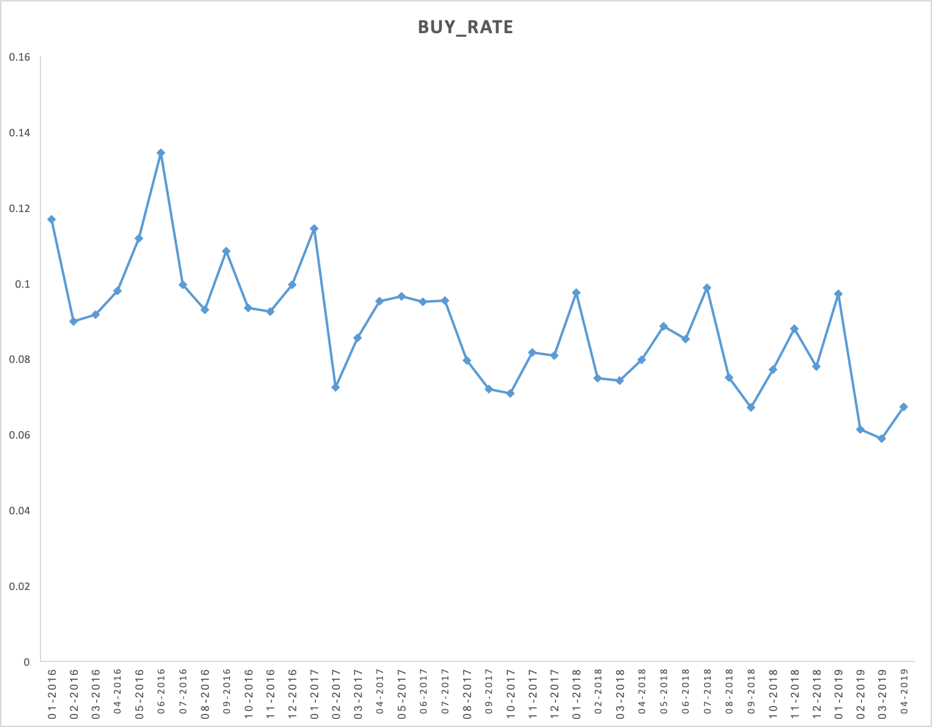
陳宜君 (B04702091)

藤田教譽 (B05705034)

陳牧忠 (P05751019)

**題目：找出關鍵少數**

我們這組希望可以“預測一個月後購買商品的會員”，第一步我們先統計每個月的“購買會員”比例，如下圖所示，每個月大約有10%的會員會購買本店商品，佔整個會員的少數，若可以找的這關鍵少數則這樣的商業價值，可以減少廣告投放的費用。也可將省下來的費用，進行重點廣告投放，增加這些“購買會員”的購買商品件數或品項。



**目標**

針對這樣的預測我們“寧可錯殺一百，也不願放過一個”，故需要提高購買會員的Recall Rate，換句話說，情願提升“不購買會員”到“購買會員”的比例，也不可以錯過一個將“購買會員”漏掉的機會。

**關鍵步驟**



如上圖，我們這組先將有效的會員篩選出，可以提升整個程式的執行效率，因為Behavior的筆數很多，經過過濾可以提升一半以上的速度，再將所有的資料數據化，再將沒有數據的會員過濾，最後將整理出來的資料丟入訓練與測試模型。

**參數定義**

首先要先定義預測和分析資料的範圍，我們定義利用3個月的資料來預測未來1個月後會員是否會購買。

“購買會員”的意思就是在有任何Order資料的會員（不論Status）。反之，則是“不購買會員”

“有效會員”則是要滿足以下的三個條件

1. 從2016/1/1開始到預測前的日期。2016/1/1是order資料的開始，所以我們訂這天當開始。
2. 在分析期間的3個月內至少要有一次的動作(至少一個特徵值大於1)
3. 有OnlineMemberID

**資料分析Part 1**

下圖是利用一些Behavior的欄位來看與“購買會員”的correlation狀況。

**資料：**

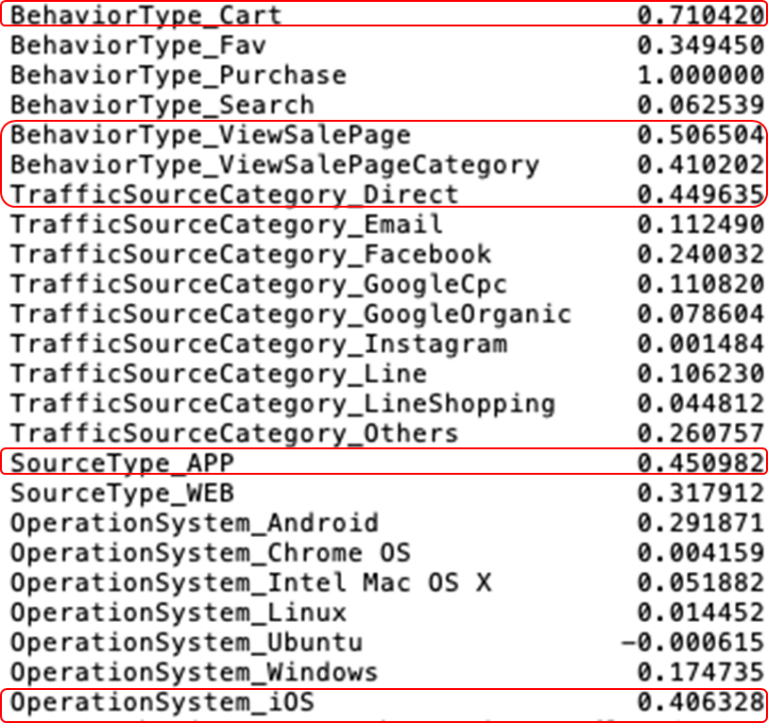
以三個月為單位，由2019/04往回切。整理出之Behavior Feature 統計數。

**有趣的發現：**

1. 客戶將商品放入購物車的次數，和購買商品的次數之相關性最為顯著。
2. 作業系統中，是否使用iOS與購買次數的相關性最高。

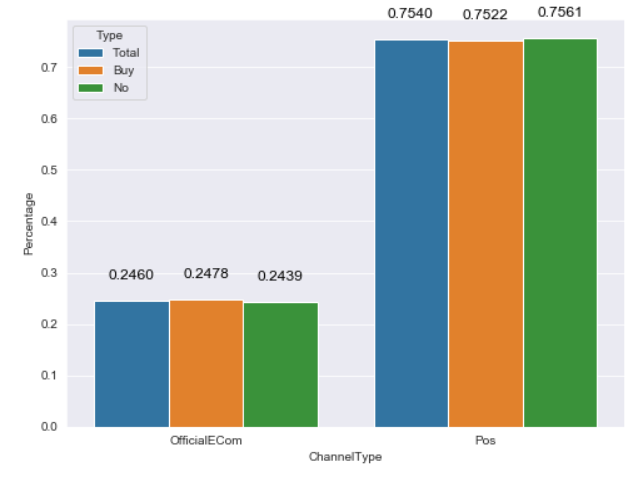
**建議：**

Cart, Viewsalepagecategory,…(correlation > 0.4)等feature與購買次數之相關性較高，在建立model做預測時，優先採用這些feature

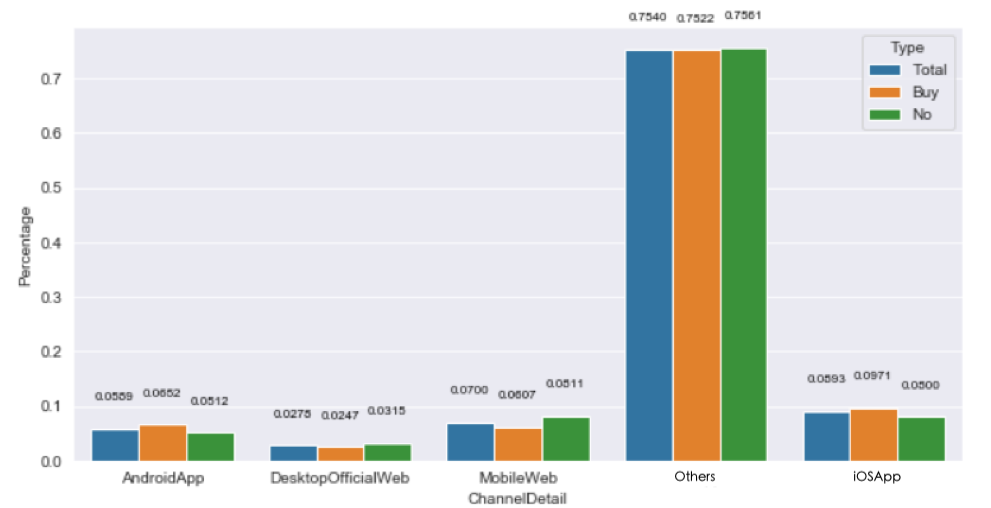


**資料分析 Part 2**

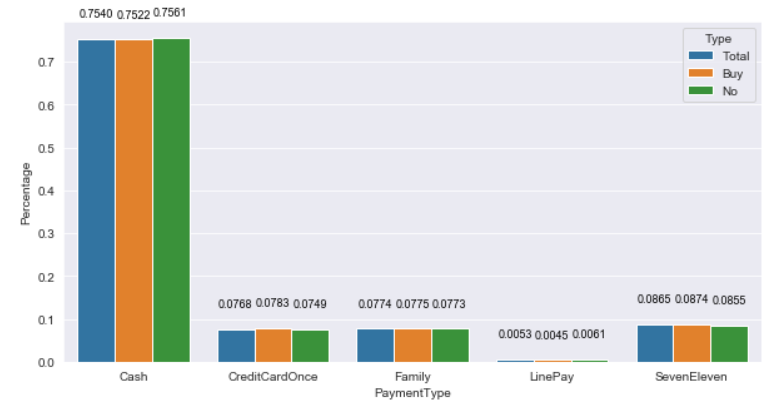
下圖是ChannelType在“購買會員”和“不購買會員”占比，並無明顯差異。



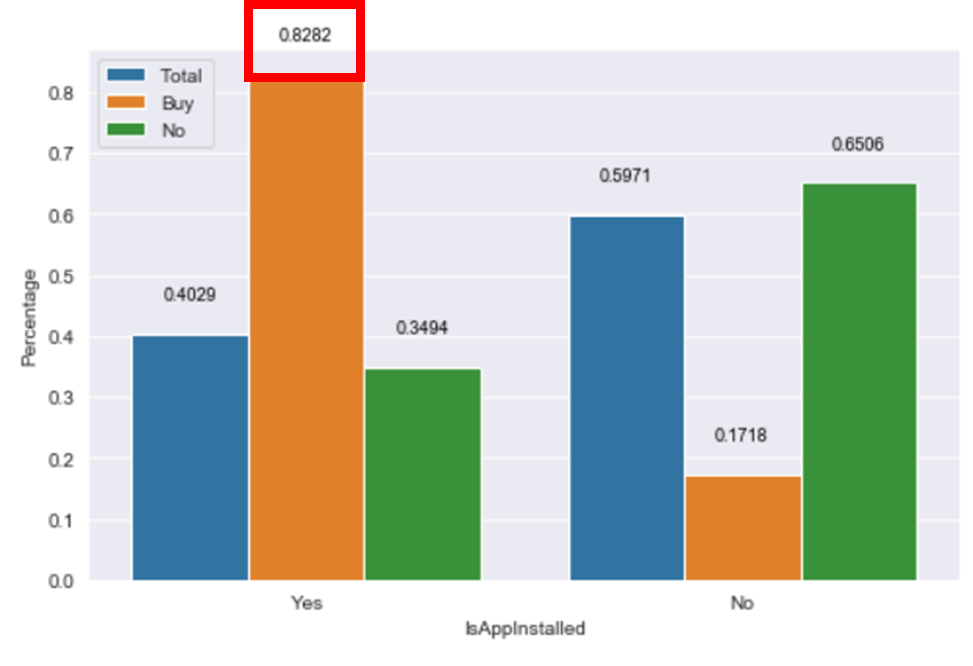
下圖是各ChannelDetail在“購買會員”和“不購買會員”占比，並無明顯差異。



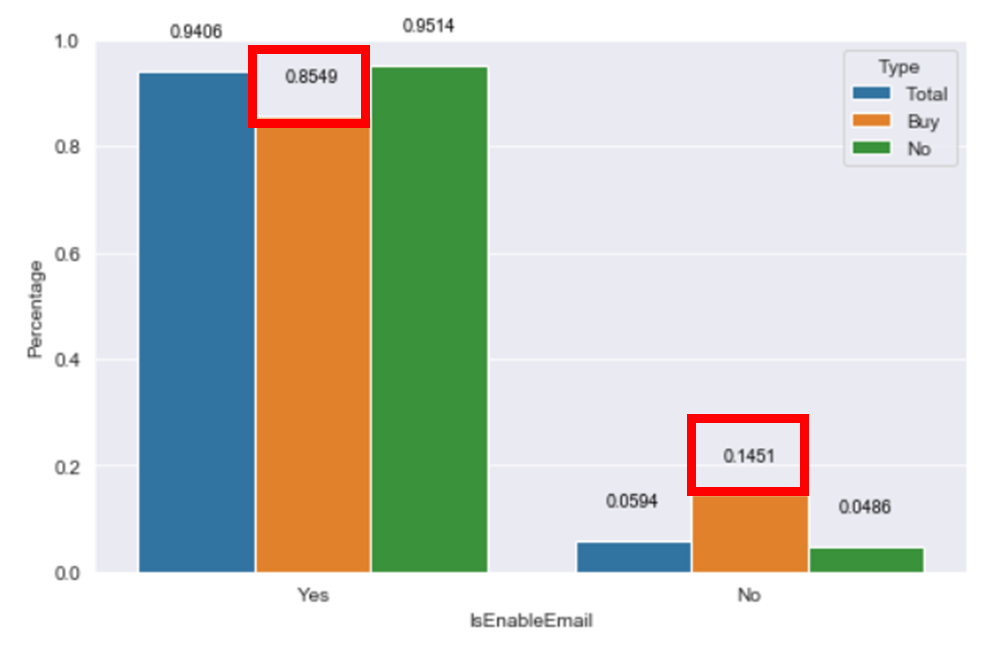
下圖是各PaymentType在“購買會員”和“不購買會員”占比，並無明顯差異。



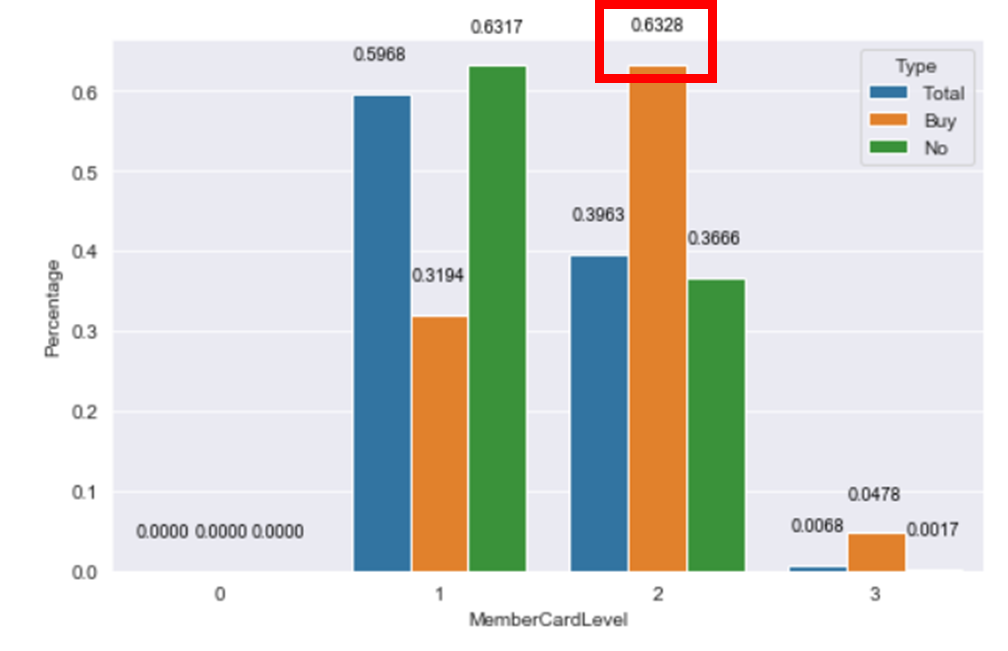
下圖是IsAppInstalled在“購買會員”和“不購買會員”占比，可以看出有裝APP的購買比例較高。



下圖是IsEnableEmail在“購買會員”和“不購買會員”占比，很有趣的是不希望收到email購買比例較高。



下圖是MemberCardLevel在“購買會員”和“不購買會員”占比，可以推斷level 2和3的購買比例較高，而level 2的“購買會員”最多。



**特徵選擇 – 69 features**

下圖是我們所選用的特徵，最後我們全部所有特徵都丟入訓練模型讓演算法自己選擇，一共69個特徵，我們將RegisterSourceTypeDef用One-Hot Encoding，其餘一些用Counter去表示，ShopPeriod和OnlineTime則是算這次和上次的差異取mean、max、min、和median。



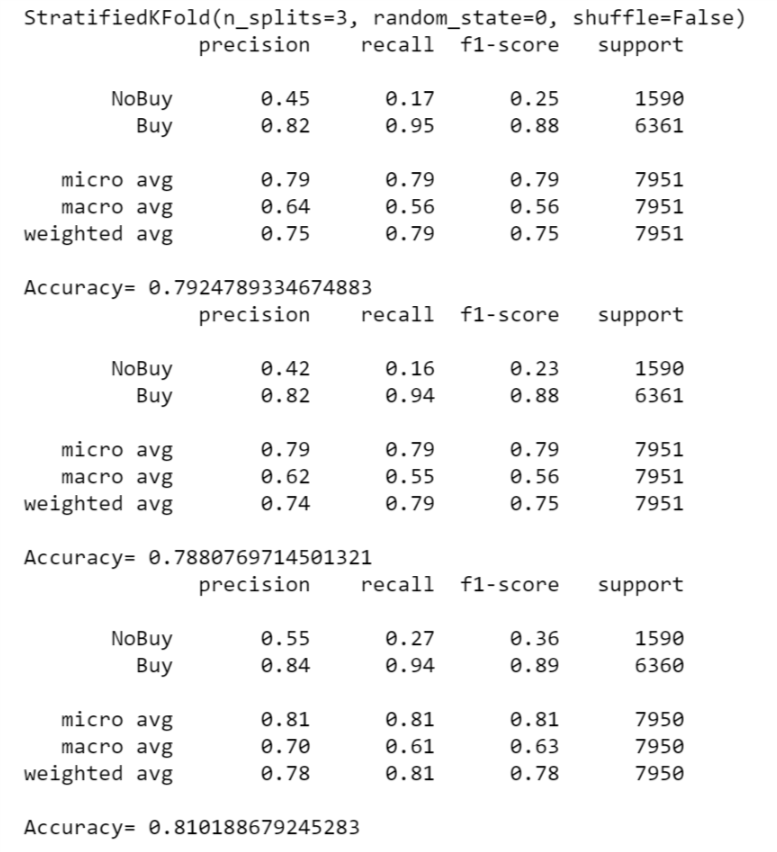
**訓練/測試場景**

如下圖所示，我們將資料切成Training set和Testing set，分析資料用3個月，用下個月貼標買或不買的會員。Training從2018/4/到2018/12/31，因為Behavior的資料啟始是從2018/3開始。而Testing的則是用最後幾個月的資料來回測判斷。



**訓練參數**

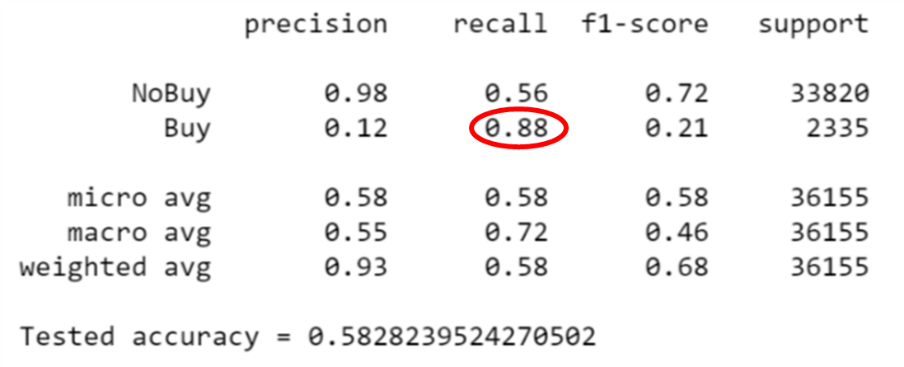
我們訓練模型是用Random Forest，K-Fold用3，所有的訓練資料經過篩選過後有23852筆，而Buy/NoBuy 的比例大約4:1，因為要讓訓練模型判斷大部分都是買的故意讓“買”的比例大過“不買”的比例。下圖則是訓練過程。



**預測結果 – 省下一半廣告費**

測試會員數在2019/4經過過濾的無效會員後有36155個而Buy/NoBuy 的比例大約是1:9如預期，我們希望可以找到關鍵的10%。

我們假設在沒有確定Lost會員的前提下，廣告投放是針對所有會員，故預測前的投放有36155個有效會員，經過預測後只剩下47%的“購買會員”，換言之就是36155\*42% =15185個會員，總共提升經濟效益有58%。如下圖，搭配約9成的Recall Rate就表示幾乎這47%裡包含了90%的“購買會員”。



**Follow-ups**

這樣的結果可以輔助NAPL模型，更精準的投放廣告給有意願購買的客人。

要如何增加準確率

1. 利用會員特徵的變化量產生新的特徵，例如這個三個月和上三個月的差異或變化量，提升準確率
2. 針對購買會員做進一步分析
   1. 可以觀察，“購買會員”的變化程度，是否這個月和上個月，甚至上上的月的會員都相同。
   2. 針對買或不買的關鍵特徵可以增加權重，在投入訓練模型裡