資料分析-消費者購買行為

楊承鑫　2024/5

1. 資料來源：Kaggle https://www.kaggle.com/datasets/sanyamgoyal401/customer-purchases-behaviour-dataset/data

2. 資料描述：

- 資料形狀：100000列12欄 (100000個消費者及每個消費者對應的12特徵)
- 12個特徵：

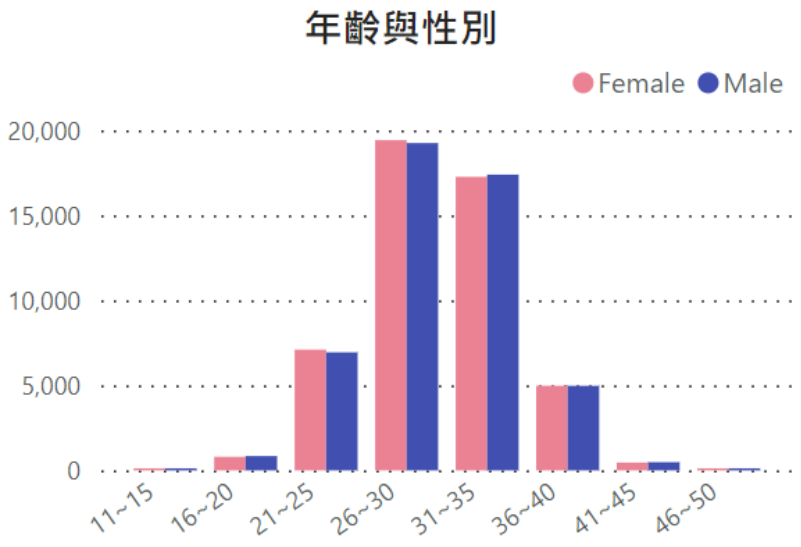
1) 身分證(id)：1, 2, ..., 100000
2) 年齡(age)
3) 性別(gender)：分為男生及女生
4) 收入(income)
5) 教育程度(education)：分為高中、學院、大學及研究所
6) 居住地區(region)：分為東、南、西、北
7) 忠誠度狀態(loyalty status)：基於購買歷史、頻率及參與品牌活動等指標，將顧客分為普通、銀牌及金牌
8) 購物頻率(purchase frequency)：分為很少、偶爾及頻繁
9) 購物金額(purchase amount)
10) 產品種類(product category)：分為書、食物、電子產品、家用產品、健康產品、服飾及美妝產品
11) 促銷使用(promotion usage)：0為沒有使用，1為有使用
12) 滿意度(satisfaction score)：1, 2, ..., 7

3. 分析目的：

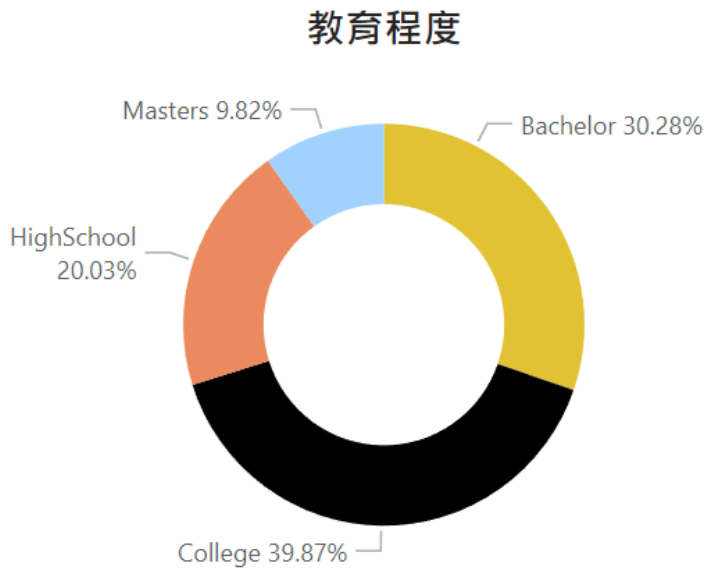
當一個新顧客來消費時，能夠透過新顧客的個人特徵(1~6)及購買特徵(8~10)預測其忠誠度狀態，也就是能夠提早知道新顧客的忠誠度給予對應的服務，例如：如果預測是金牌會員，或許可以優先處理他訂單、投訴或是給予定期優惠或獎勵，以提早穩固客戶關係。

4. 探索式資料分析

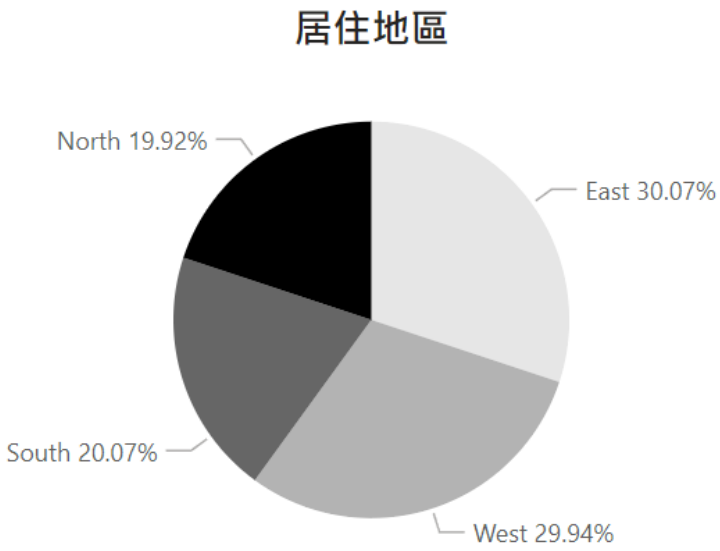
- 資料無遺失值及離群值
- 數據統計圖表：



26~30歲的顧客最多。每個年齡區間男女比例相近。

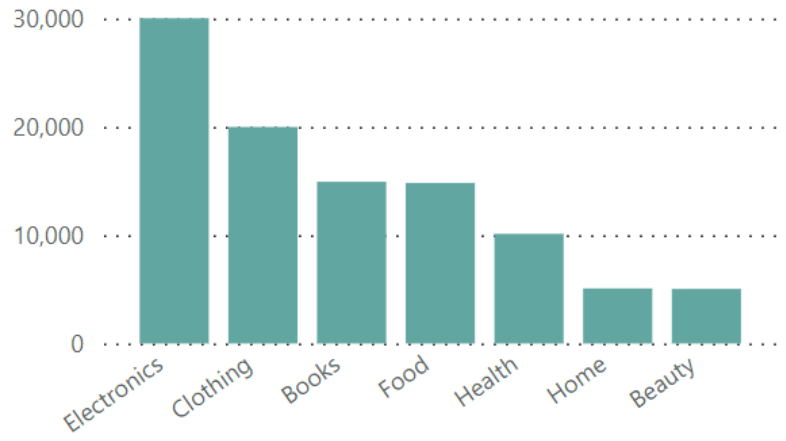


顧客教育程度大部分是學院與大學。



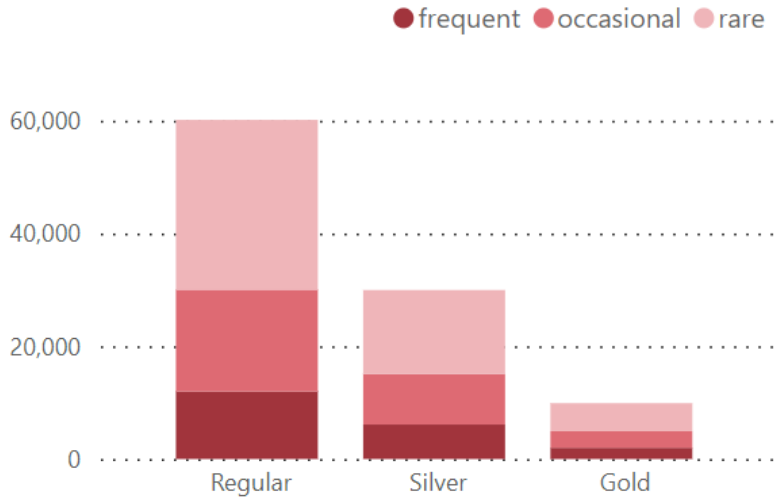
顧客大多來自東部與西部

產品種類



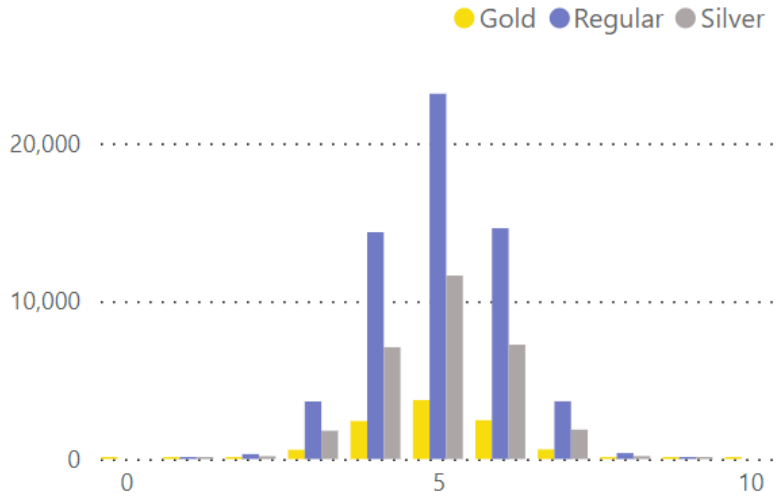
顧客大多購買電子產品及服飾。

忠誠度狀態與購物頻率



在三種忠誠度狀態的顧客間，購物頻率的比例並沒有顯著差異。並不會有金牌會員購物頻率較高的現象。

滿意度與忠誠度狀態

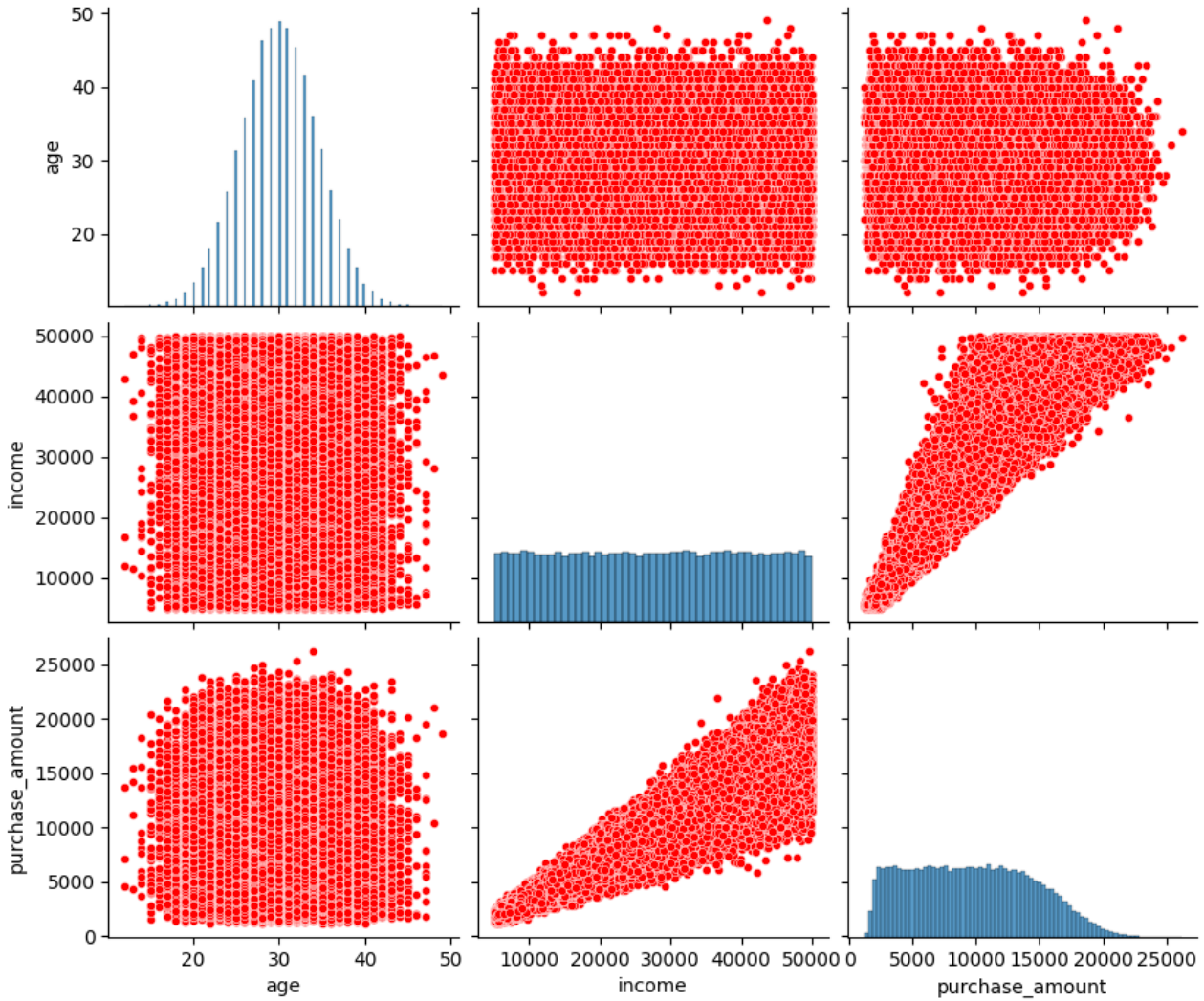


顧客滿意度呈現鐘型分配，各個滿意度間顧客的忠誠度狀態比例無明顯差異。

相關係數矩陣(Correlation Coefficient Matrix)

	age	income	purchase_amount
age	1.0000	-0.0024	-0.0021
income	-0.0024	1.0000	0.9484
purchase_amount	-0.0021	0.9484	1.0000

散點圖矩陣(Pairs Plots)



- 年齡(在顯著水準為0.05時)不拒絕服從常態分佈(p-value=0.79)，平均值為30，變異數為20
- 顧客中收入最低為5000，最高為50000，平均值為27516 ，分布均勻，無趨勢。
- 購物金額最低為1118，最高為26024，平均值為9635，顧客數在購物金額大於13000時呈現遞減狀態。
- 年齡與收入和年齡與購物金額皆呈現零相關性。
- 收入和購物金額相關係數為0.95 ，為高度正相關性。

5. 資料預處理：

- 刪除對預測無幫助的欄位：身分證
- 對類別數等於 2 且類別間無高低之分的欄位進行Label encoding：年齡
- 對類別數超過 2 且類別間無高低之分的欄位進行One-hot encoding：居住地區、產品種類
- 對類別間有高低之分的欄位進行Ordinal encoding：教育程度、購物頻率

處理後資料如下：

age	income	purchase_amo	promotion_us	satisfaction_s	gender	education	North	East	West	South	purchase_fre	Beauty	Clothing	Health	Home	Electronics	Food	Books	loyalty_status
24	47773	21794	0	5	0	0	1	0	0	0	0	0	0	0	0	0	0	1	Regular
27	40682	18249	0	6	1	2	0	1	0	0	2	0	0	0	0	0	0	1	Gold
27	19154	5819	0	5	1	1	0	1	0	0	1	0	1	0	0	0	0	0	Regular
28	24666	8779	0	6	0	0	1	0	0	0	0	0	0	0	0	0	1	0	Regular
28	35748	12901	1	3	1	2	1	0	0	0	0	0	0	0	0	0	0	1	Silver
29	15317	4557	1	6	1	3	0	0	1	0	0	0	1	0	0	0	0	0	Regular
30	11568	4098	0	7	1	0	0	0	0	1	2	0	0	0	0	0	1	0	Regular
30	19034	5579	1	5	0	2	0	1	0	0	1	0	0	0	0	0	0	1	Regular
31	46952	19685	1	5	0	1	1	0	0	0	1	0	1	0	0	0	0	0	Regular
32	8265	3293	0	7	0	2	0	0	0	1	2	0	1	0	0	0	0	0	Silver
32	40044	13608	0	5	0	2	1	0	0	0	0	0	0	0	0	1	0	0	Silver
32	6735	2450	1	5	1	1	0	1	0	0	1	0	1	0	0	0	0	0	Silver
35	43896	16158	1	6	1	0	0	0	0	1	0	0	0	0	1	0	0	0	Regular
37	38849	11822	0	6	1	2	0	0	1	0	0	0	1	0	0	0	0	0	Silver
38	7347	2822	0	5	1	2	0	0	0	1	1	0	0	0	0	1	0	0	Silver

6. 預測方法：

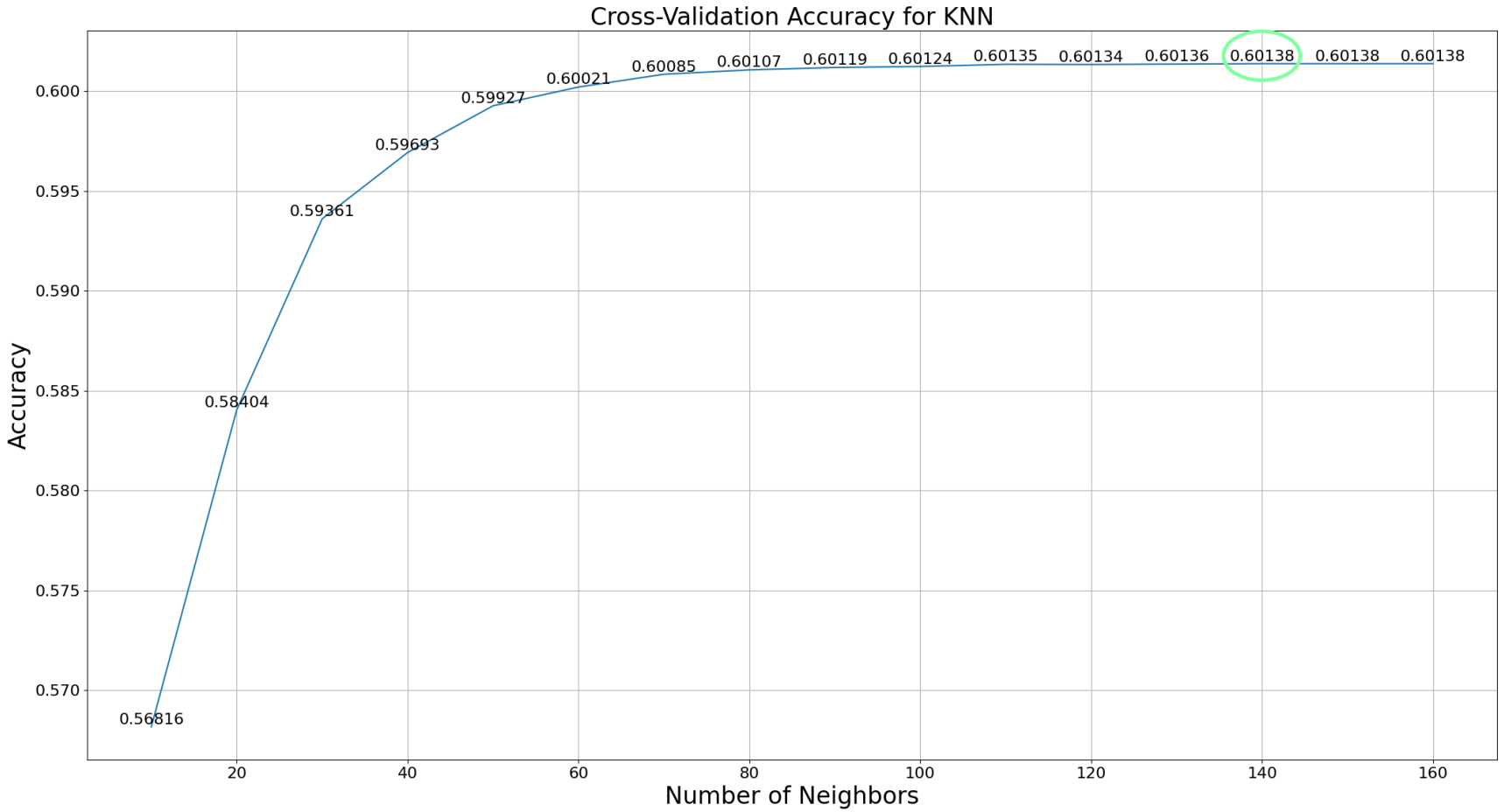
- K Nearest Neighbors (KNN)
- 集成學習：eXtreme Gradient Boost (XGBoost)
- 深度學習：Neural Network

7. 模型訓練：

- 將資料分為Y(忠誠度狀態)及X(剩下的資料)
- 對X進行標準化
- 利用10折交叉驗證(10-Fold Cross-Validation)的準確率(Accuracy)和演算法的執行時間，找出最佳演算法及對應的最佳超參數(Hyper Parameters)

方法一. K Nearest Neighbors: 離預測點最近的K個訓練點中，哪個類別最多，此預測點就分到那個類別。

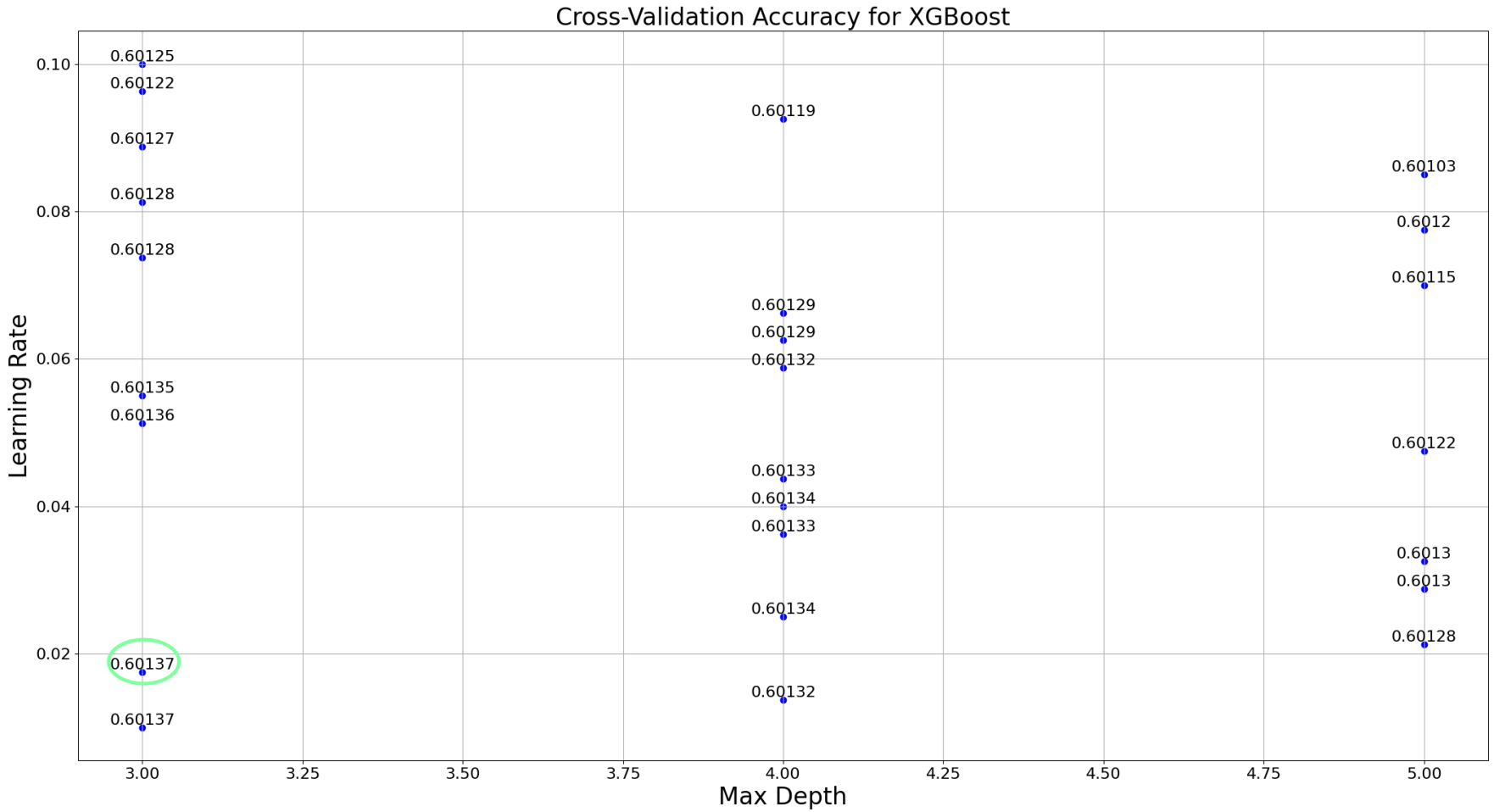
調節參數：鄰近點(Neighbors)個數(10~160)



最佳鄰近點個數為140，準確率為0.60138

方法二. XGBoost: 結合多個決策樹形成一個強大的預測模型。用當前的樹修正上一棵樹的殘差(預測與實際的差距)，最有將所有樹加起來得到預測結果。此方法在結構化數據中表現優異。

調節參數(隨機值選取)：學習率(Learning rate)(0.05~0.1)、樹的最大深度(3, 4, 5)

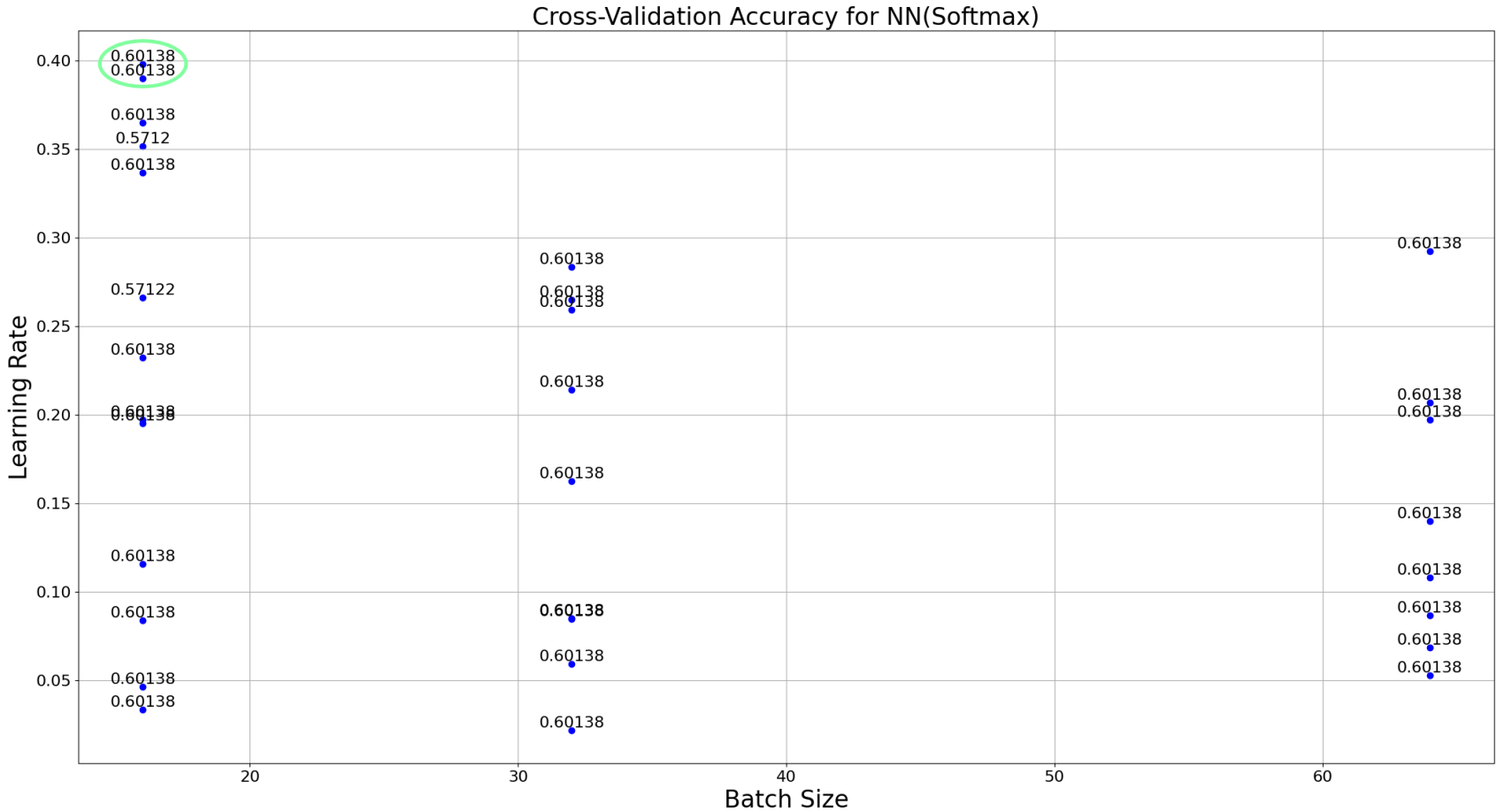


同樣準確率，選擇執行速度最快的，最佳(學習率, 樹深度)為(0.17, 3)，準確率為0.60137

方法三. Neural Network (Softmax): 透過多個神經元多層連結完成複雜且非線性的分類器。

設定： * 兩層隱藏層(Hidden layer)
* 神經元(Neuron)各數分別為8及8，其中Dropout rate為0.2, 激勵函數(Activation function)為ReLU
* Adam演算法的參數使用的預設
* Epoch為3

調節參數(隨機值選取)：學習率(0.01~0.4)、小批次大小(Mini-batch size) (16, 32, 64)



同樣準確率，選擇執行速度最快的，最佳(學習率, 小批次大小)為(0.4, 16)

方法比較.

方法	CV準確率	執行時間
K Nearest Neighbors	60.138%	0.13s
XGBoost	60.137%	0.42s
Neural Network	60.138%	6.83s

三個方法在測試資料上準確率皆約60.138%，然而K Nearest Neighbors執行時間低於其他兩種方法。

8. 結論：

當100個新顧客來購物且留下資訊時，我們可以成功預測約60.1個顧客的忠誠度狀態，提早給予對應的服務，建立牢固的客戶關係。

報告結束，謝謝!