資料分析-消費者購買行為

楊承鑫　2024

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：

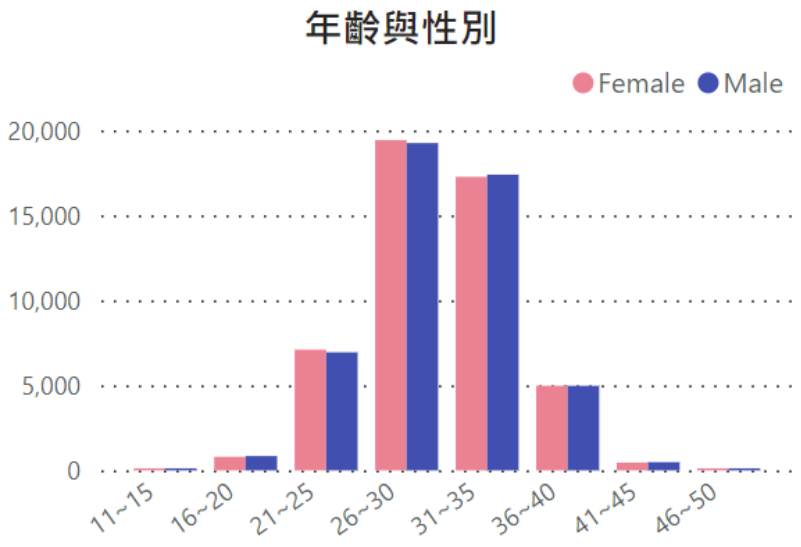
1) 身分證(id)：1, 2, ..., 100000
2) 年齡(age)
3) 性別(gender)：分為男生及女生
4) 收入(income)
5) 教育程度(education)：分為高中、學院、大學及研究所
6) 居住地區(region)：分為東、南、西、北
7) 忠誠度狀態(loyalty status)：基於購買歷史、頻率及參與品牌活動等指標，將顧客分為普通、銀牌及金牌
8) 購物頻率(purchase frequency)：分為很少、偶爾及頻繁
9) 購物金額(purchase amount)
10) 產品種類(product category)：分為書、食物、電子產品、家用產品、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. 分析目的及方法：

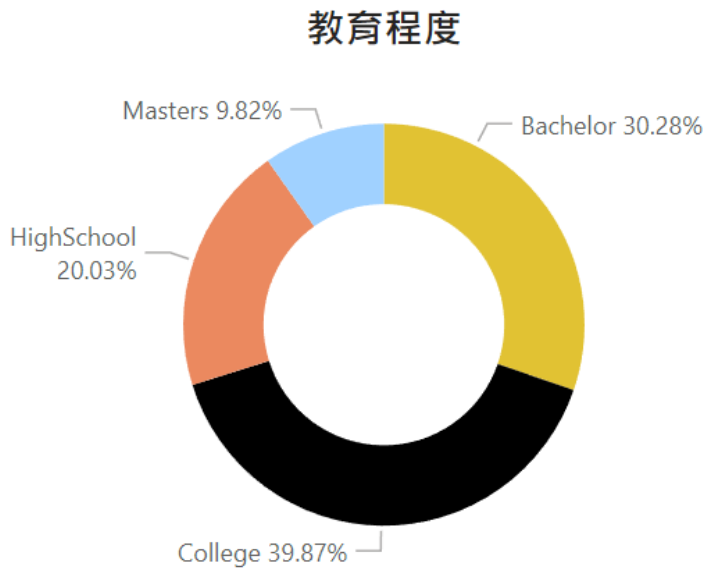
當一個新顧客來消費時，能夠透過新顧客的個人特徵(1~6)及購買特徵(8~10)預測其忠誠度狀態，也就是能夠提早知道新顧客的忠誠度給予對應的服務，例如：如果預測是金牌會員，或許可以優先處理他訂單、投訴或是給予定期優惠或獎勵，以提早穩固客戶關係。使用方法為：KNN、XGBoost

4. 探索式資料分析

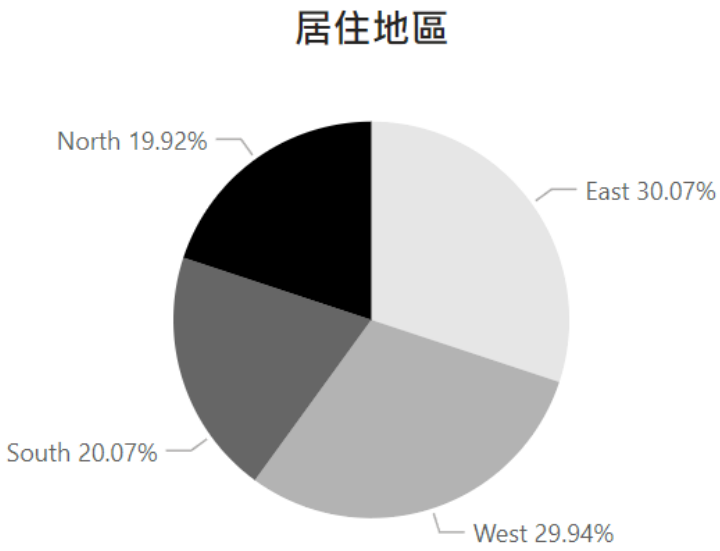
- 數據統計圖表：



26~30歲的顧客最多。每個年齡區間男女比例相近。

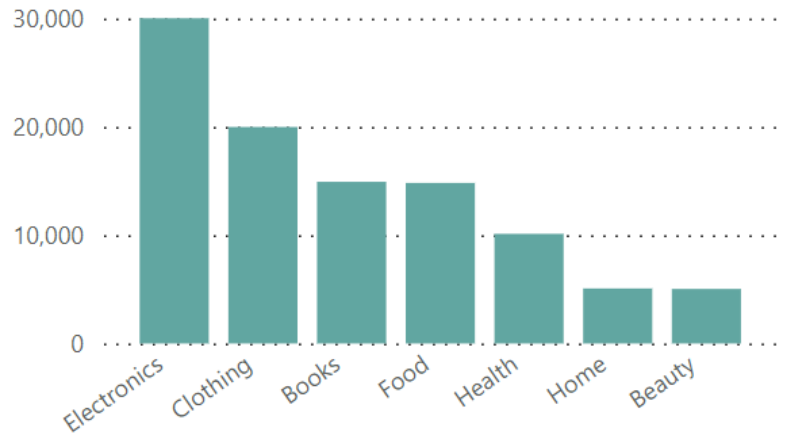


顧客教育程度大部分是學院與大學。



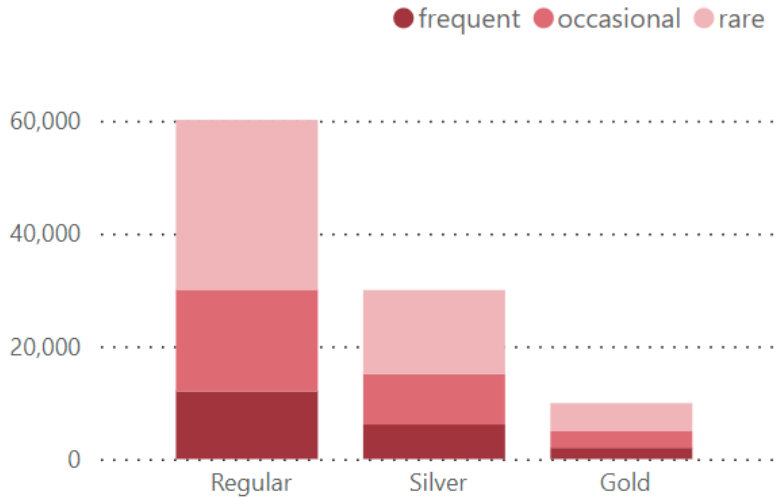
顧客大多來自東部與西部

產品種類



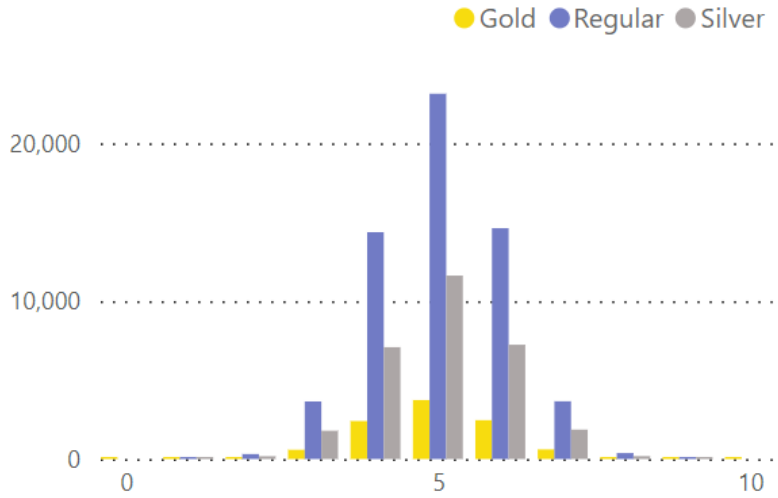
顧客大多購買電子產品及服飾。

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

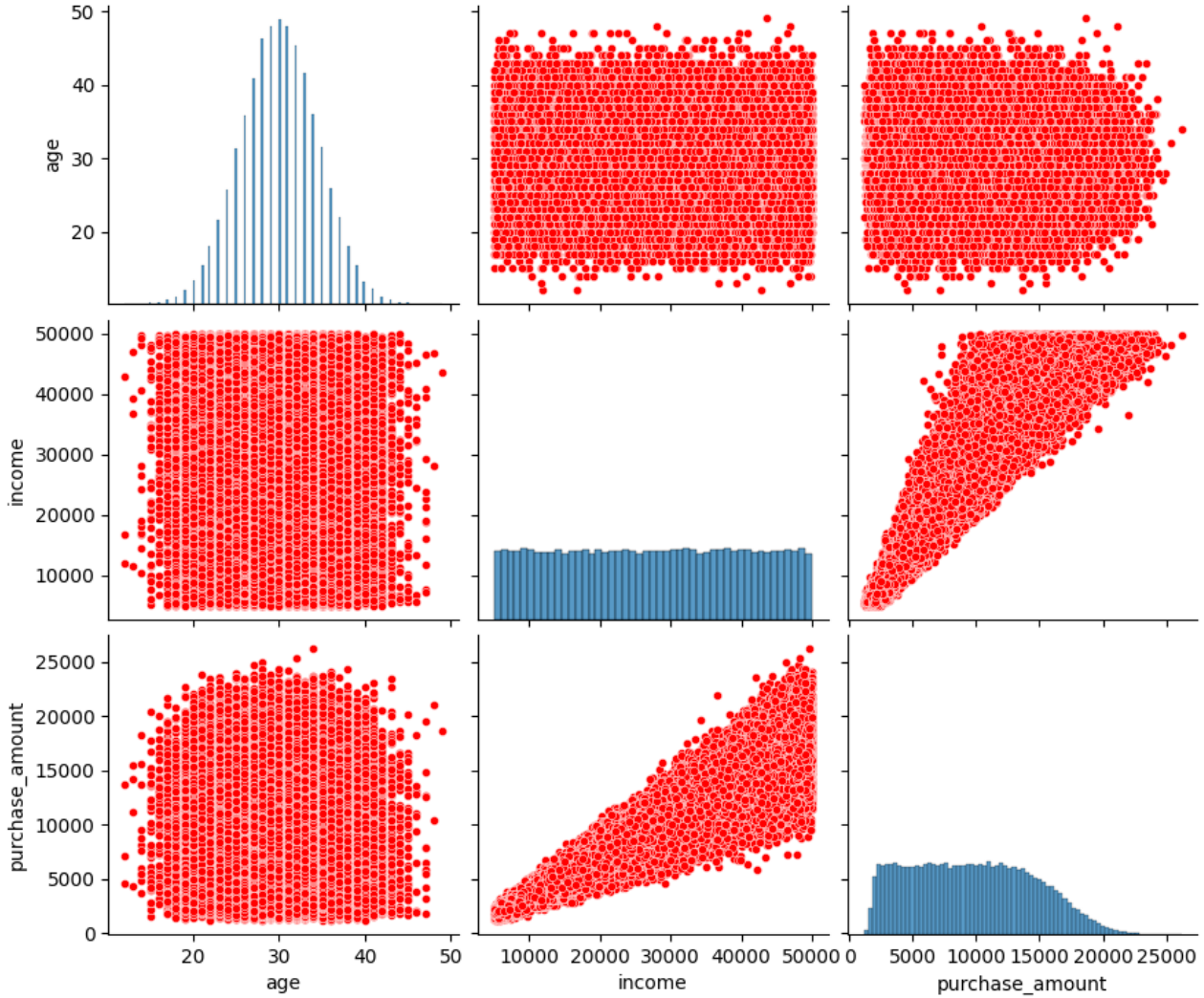


顧客滿意度呈現鐘型分配，各個滿意度間顧客的忠誠度狀態比例無明顯差異。

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：

- 資料無遺失值及異常值
- 刪除對預測無幫助的欄位：身分證

6. 特徵工程：

K Nearest Neighbors

- 類別間無高低之分的欄位進行頻率編碼(Frequency Encoding)：居住地區、產品種類
- 類別間有高低之分的欄位進行序號編碼(Ordinal Encoding) ：教育程度、購物頻率、性別(只有兩個類別)
- 特徵選取：使用遞迴特徵消除(Recursive Feature Elimination) (補充：通常特徵越少對於KNN的表現越好)
- 標準化(特徵尺度容易影響結果)

XGBoost

- 類別間無高低之分的欄位進行頻率編碼：居住地區、產品種類
- 類別間有高低之分的欄位進行序號編碼：教育程度、購物頻率、性別(只有兩個類別)
- 特徵選取：根據XGBoost特徵貢獻性(包含增益(Gain)、覆蓋率(Cover)、頻率(Frequency))選取特徵
- 創建特徵：考慮二次多項式特徵

7. 模型訓練：

- 將資料分為目標變數Y及特徵矩陣X
- 取90%資料為訓練資料集(Training Set)， 10%資料為測試資料集(Test Set)

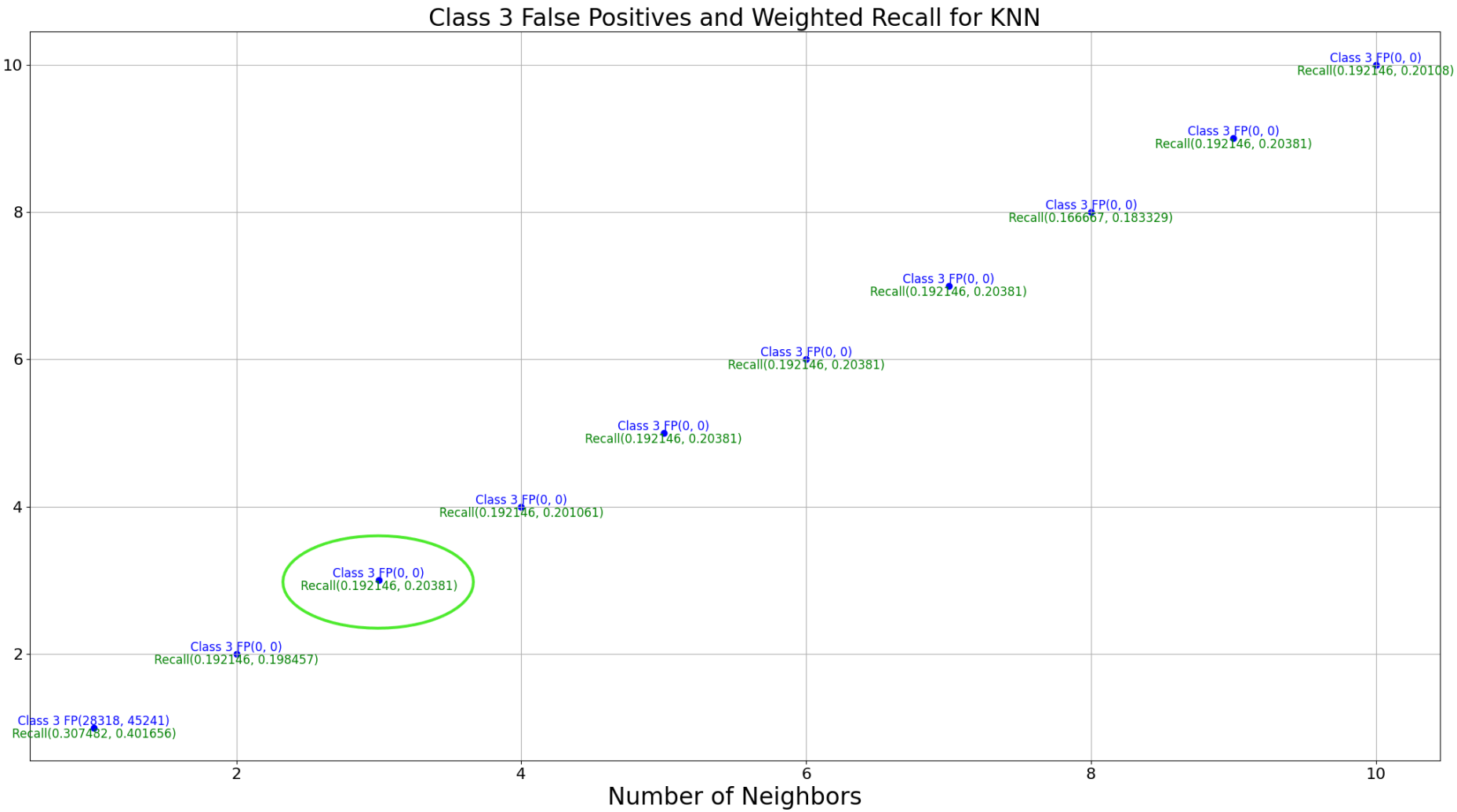
- 參數調適準則(Metrics)：
  最佳化(Optimizing)：交叉驗證(Cross-Validation)的權重召回率(Weighted Recall)
  滿足 (Satisficing)：交叉驗證(Cross-Validation)的金牌會員假正(False Positive)數量 ≤ 10
  利用準則找出演算法(Algorithm) 對應的最佳超參數(Hyper Parameters)

- 演算法比較準則：
  最佳化(Optimizing)：真實(True)的權重召回率

K Nearest Neighbors

超參數：鄰居數(Number of Neighbers)

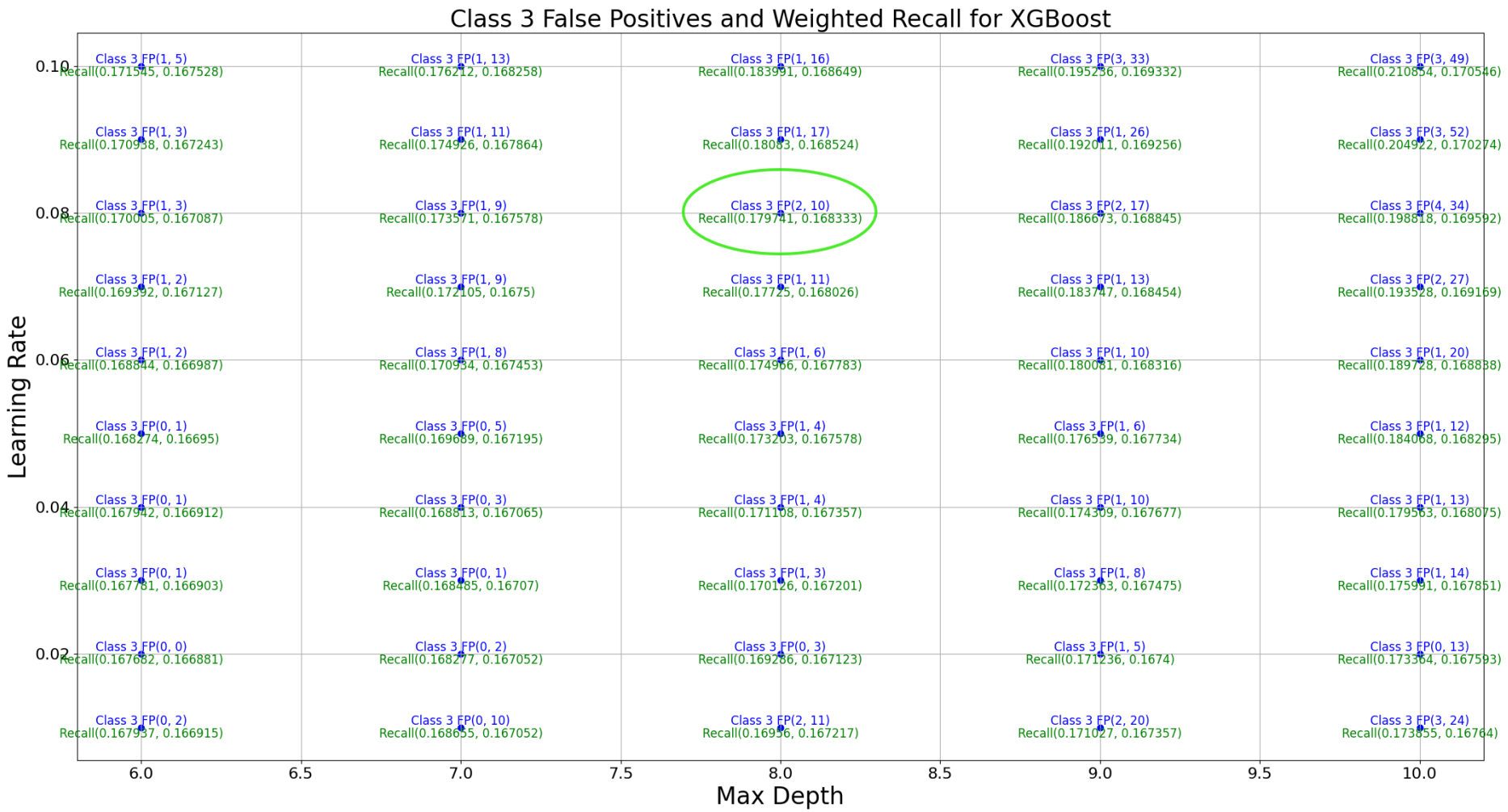
(In-sample, CV true) 表現如下圖：最佳鄰居數為3



XGBoost

超參數：學習率(Learning Rate)、樹的最大深度
搜尋方法：網格搜尋(Grid Search)

(In-sample, CV true) 表現如下圖：最佳(學習率, 最大深度)為(0.08, 8)



演算法方法比較

方法	真實權重招回率
K Nearest Neighbors	19.35%
XGBoost	16.71%

K Nearest Neighbors對於此分析目的的表現優於XGBoost

報告結束，謝謝!