建立信用卡詐欺的二分判斷系統

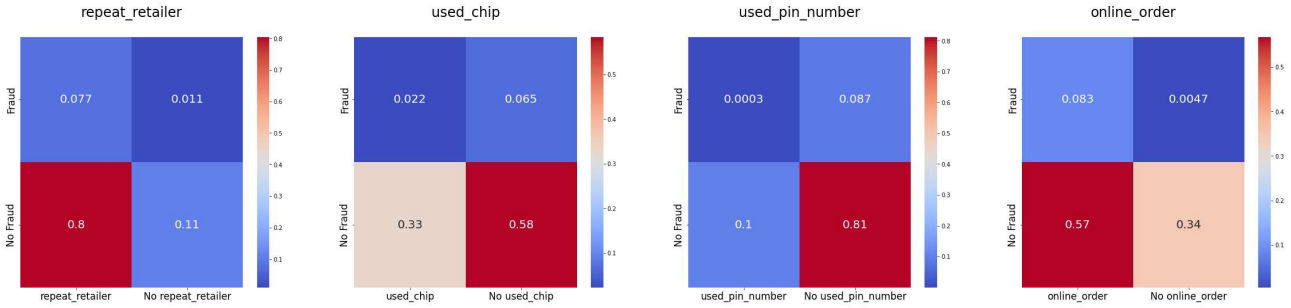
本次研究的資料集——「Credit Card Fraud」，取自 Kaggle 網站(https://reurl.cc/j191kq)，記錄了一百萬筆信用卡交易的各項性質，並註明每筆交易是否為詐欺。如能透過交易的性質判斷詐欺的可能性，有關單位便能提前展開調查。而本文目的即是藉由該資料集建立判別信用卡詐欺的分類系統。

首先簡介資料集，它並未區分訓練集與測試集，共一百萬筆資料，八個欄位，分別的意義如下：

1. distance_from_home - 與住家的距離。
2. distance_from_last_transaction - 與上次交易位置的距離
3. ratio_to_median_purchase_price - 與中位交易金額的比值
4. repeat_retailer - 是否來自同樣零售商
5. used_chip - 是否使用晶片
6. used_pin_number - 是否使用 PIN 碼
7. online_order - 是否為網路訂單
8. fraud - 是否為詐欺

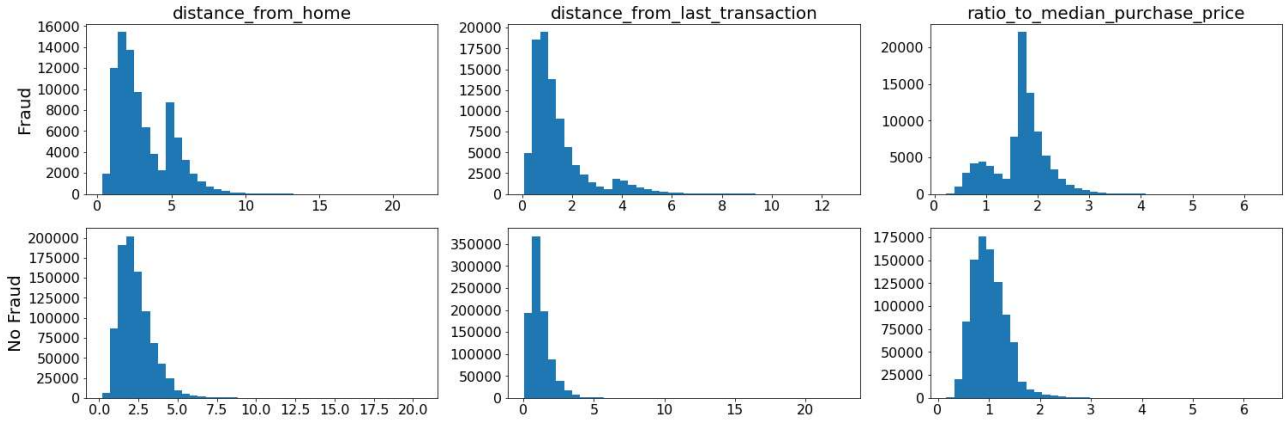
從欄位意義可以明白，任務即是透過欄位 1~7 預測欄位 8 的詐欺與否。而欄位 4~8 為二元資料，其餘為連續性資料。

由於詐欺案例應占少數，經查看約佔 8.74%，單就案例數量而言，生成的分類器未必能準確判斷詐欺。於是查看詐欺與非詐欺的狀況下，各欄位的分布是否對於詐欺有所偏向。對於二元欄位，將其依詐欺與否、該欄位的二元與否，共分四類，將這四類於欄位中的占比繪製如下。



從 repeat_retailer、used_pin_number、 online_order 的繪圖可以看出，一旦發生詐欺，通常來自同一零售商，幾乎都不使用 PIN 碼且為網路訂單。

接著檢視連續性欄位，同樣分為詐欺與非詐欺的狀況。首先查看分布，由於離群值較多，繪圖之前先取了立方根。



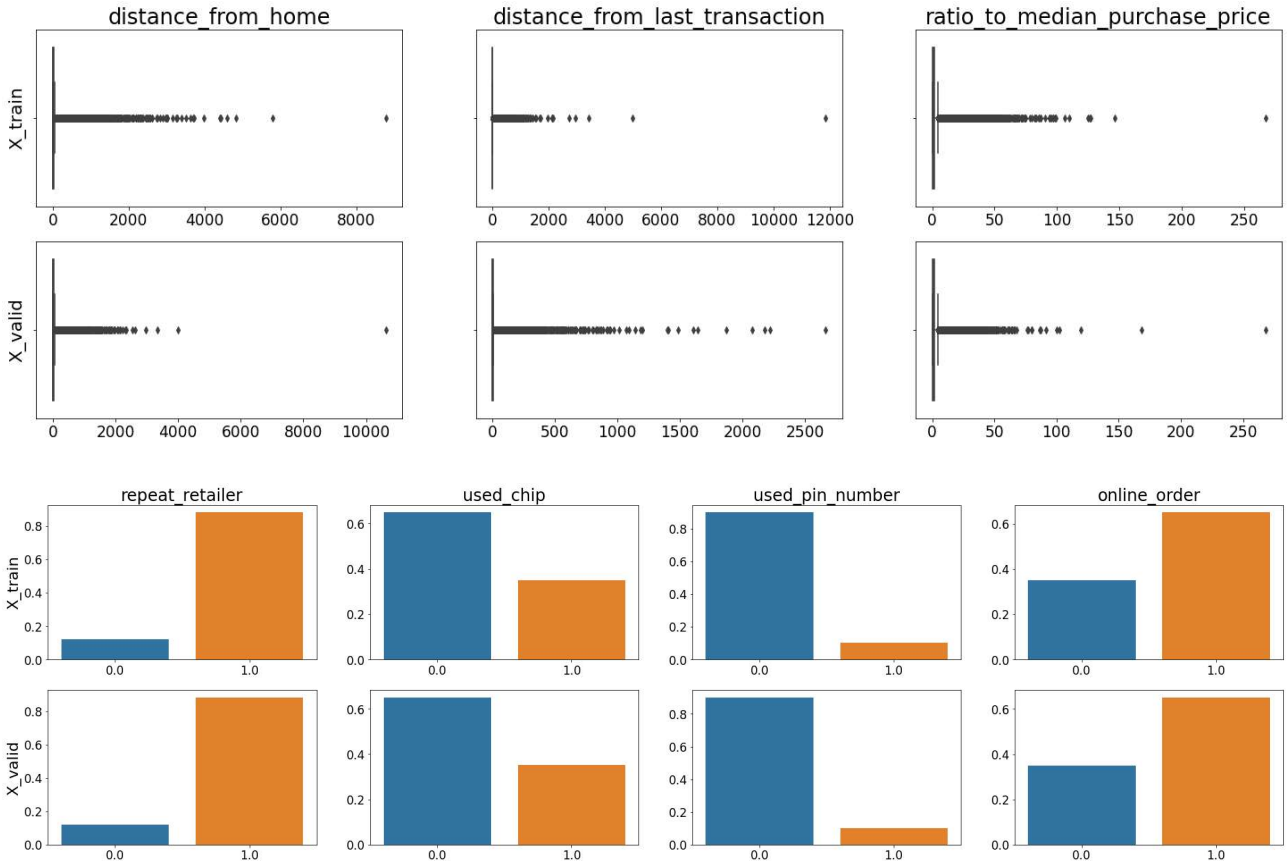
從 distance_from_home、distance_from_last_transaction 的繪圖可以見得，非詐欺的狀況下，交易離住家的距離、與上次交易位置的距離是遞減的，亦即交易位置不應離住家太遠。然而詐欺的狀況下，遠

離一段距離後交易卻又興盛起來。ratio_to_median_purchase_price 也有類似的遞減狀況，而詐欺時尤為誇張，從大量的高額爆刷可以感受到人類的惡意。綜上所述，詐欺通常伴隨幾種狀況：

1. repeat_retailer = 1
2. used_pin_number = 0
3. online_order = 1
4. Large distance_from_home
5. Large distance_from_last_transaction
6. Large ratio_to_median_purchase_price

接著構建分類器，並將分類結果與上述結論比對。

將資料集以 3:1 分拆為訓練集、驗證集，由連續型欄位的箱型圖確認兩者有相近的統計量，而二元欄位數量也有相近的比例。



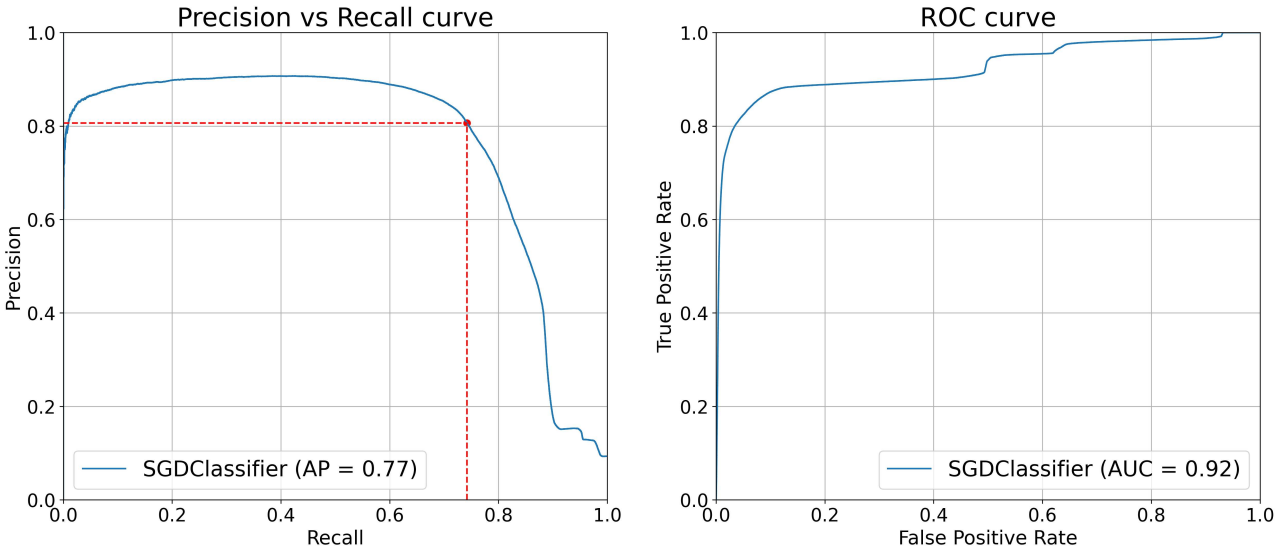
該分類器必須有高 recall 以找出盡可能多的詐欺案例，且有足夠的 precision 以防太多誤判；由於兩者難以兼具，額外使用 f1-score 作為評分標準。關於生成方法，因測試集有七十五萬筆數據，考慮計算時間，使用 sklearn 的 LogisticRegression(邏輯回歸)與 SGDClassifier(隨機梯度分類)，其中後者擬合線性 SVM。訓練過程取 K=3 做 K-fold 交叉驗證，並以 f1-score 作評分標準。

	Estimator	score1	score2	score3	Time duration
0	LogisticRegression	0.721	0.681	0.702	15.9 s
1	SGDClassifier	0.758	0.775	0.775	118.0 s

上表顯示 SGDClassifier 約耗 2 分鐘完成交叉驗證(相對地 LogisticRegression 僅花 16 秒)，但每次驗證的 f1-score 皆較優，故選此分類器作後續討論。接著進行參數調整，由於類別數量懸殊，可在損失函數中給予類別權重；並且為了避免過擬，可添加正則化因子。以下比較不同權重與 L2 正則化因子的得分狀況，以下表格為交叉驗證的平均 f1-score。

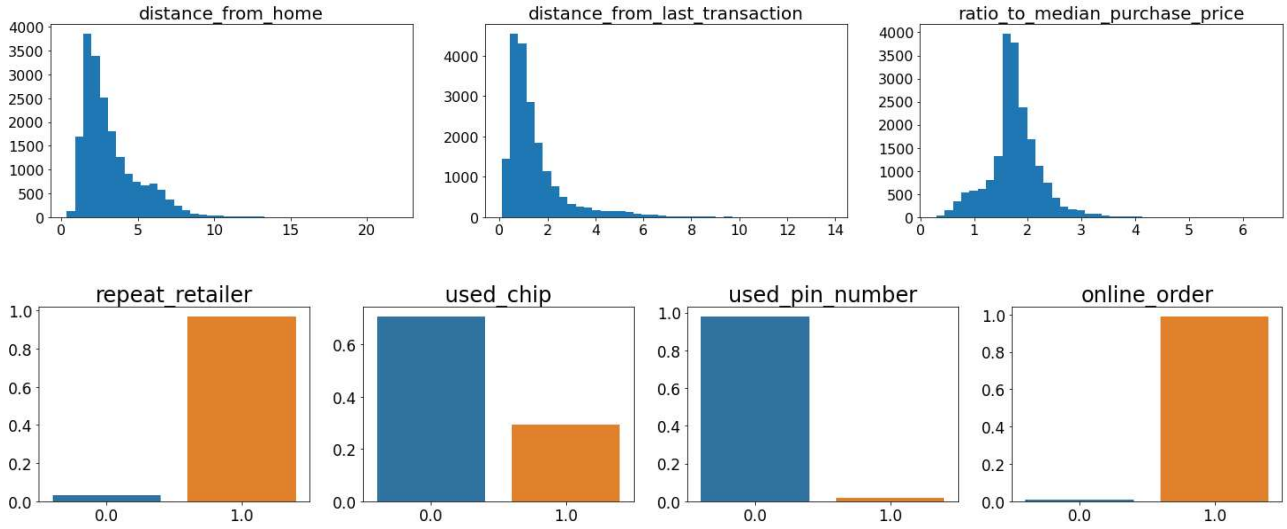
	{0: 1.0, 1: 0.5}	None	{0: 1.0, 1: 5.0}	{0: 1.0, 1: 10.0}
5e-05	0.766	0.75	0.758	0.734
0.0001	0.525	0.763	0.755	0.719
0.0003	0.691	0.753	0.715	0.697
0.0005	0.729	0.734	0.714	0.683

該表的欄位為權重組合(None 表示無加權)，列方向為 L2 正則化因子係數。可以看到無加權的狀況下，分類器對於不同因子係數的表現較高且穩定，可採用 [0.0001 , None] 的組合。經查看該分類器的 precision 為 0.832，而 recall 為 0.734，並繪出 Precision vs Recall curve 與 ROC curve (Receiver Operating Characteristic curve) 如下。



左圖看到 precision、recall 隨分類器閾值增加而有所消長的情形。紅點為當前分類器坐落位置，也位於曲線拐點附近，此處同時得到不錯的 precision、recall。右圖也是對閾值的變化，顯示在極低的 False Positive Rate 得到可觀的 True Positive Rate (約為 0.8)。

進入驗證階段，代入驗證集得 precision 為 0.827，而 recall 為 0.734，即判斷詐欺有 82.7%的正確率，且能找出所有詐欺案件中的 73.4%。以下繪出分類器預測為詐欺案件的欄位分布。



如此分布符合先前對詐欺案件的欄位總結，不過 distance_from_home 與 distance_from_last_transaction 的遞減後再起現象，未若預想地那樣明顯，僅能些微看出。除了分類器地效能外，這也與拆分驗證集時是否納入足夠的欄位離群值與實際詐欺案例有關。