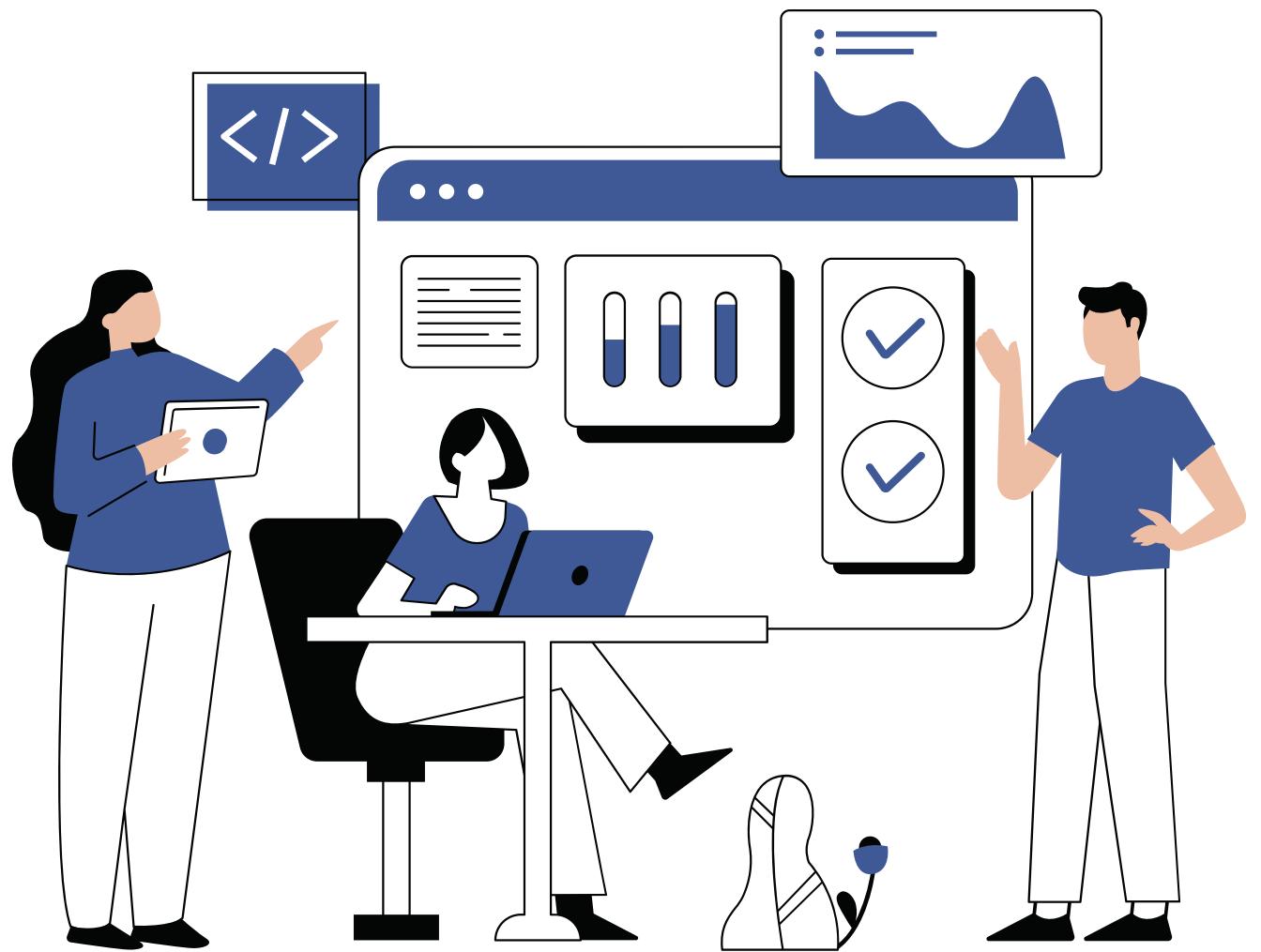


# Credit-G專題

Presented by 王珩 2023/05/09



“

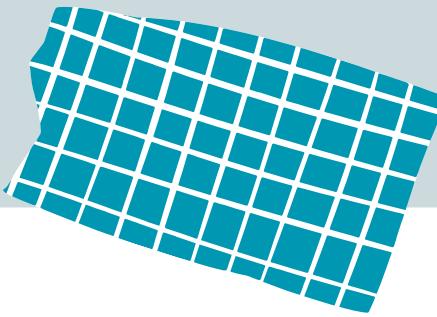
## CREDIT-G DATASET

這是OpenML上的一個數據集，它包含了銀行客戶的個人和財務信息，裡面有20個輸入特徵例如年齡性別、信用卡餘額、欠款金額、付款延遲次數和時間。輸出變量是一個二元變量，用於表示客戶是否違約（1表示違約，0表示未違約）。

我們的目標是通過輸入特徵來預測客戶是否會違約。當銀行收到貸款申請時，銀行必需根據申請人的資料決定是否通過貸款。



# 1. 定義問題



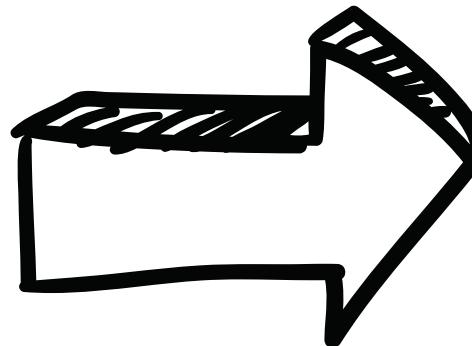
Credit-G數據分析的目的是為了評估借款人可能違約的風險，透過模型分析過去的歷史信用記錄、財務狀況相關信息，可以對借款人的信用風險進行量化評估，幫助金融機構評估借款人的信用風險，避免不必要的損失，從而更好做出相應的貸款決策。



## 2. 整理資料

將輸入特徵進行特徵提取，並進行資料清理，以確保資料的可用性和提供給模型進行預測分析。

<b>id</b>	<b>checking_status</b>	<b>credit_amount</b>
0	'<0'	1169
1	'0<=X<200'	1169
2	'no checking'	2096



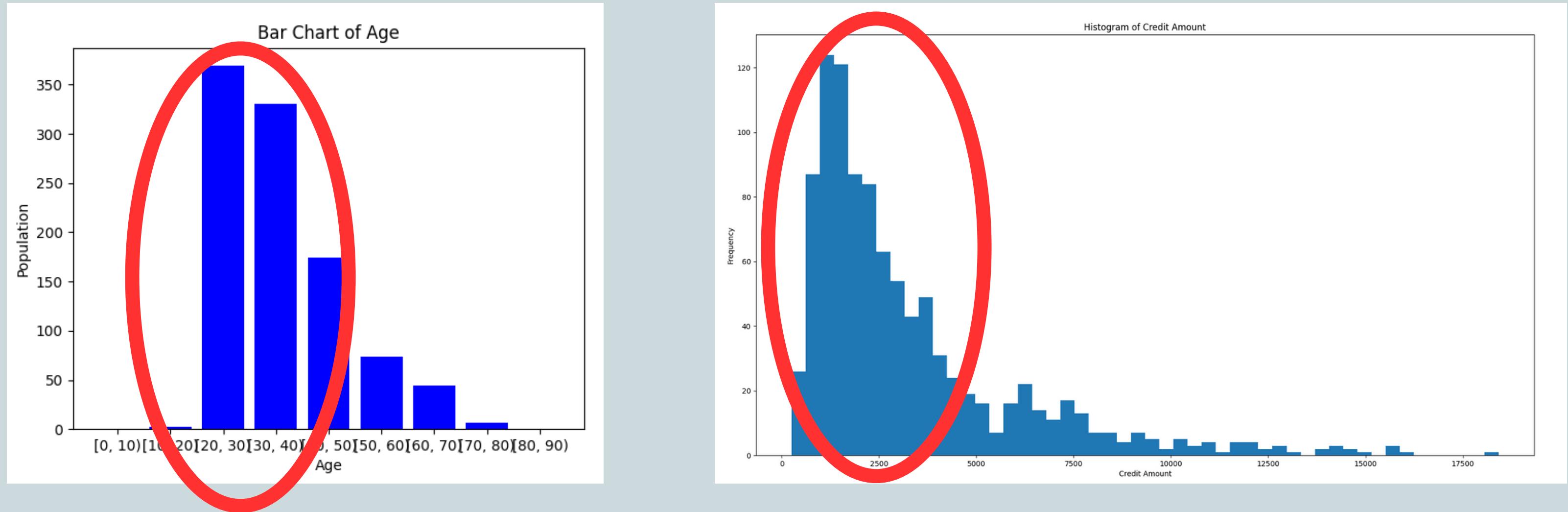
<b>id</b>	<b>checking_status</b>	<b>credit_amount</b>
0	1	1169
1	0	1169
2	2	2096

Before

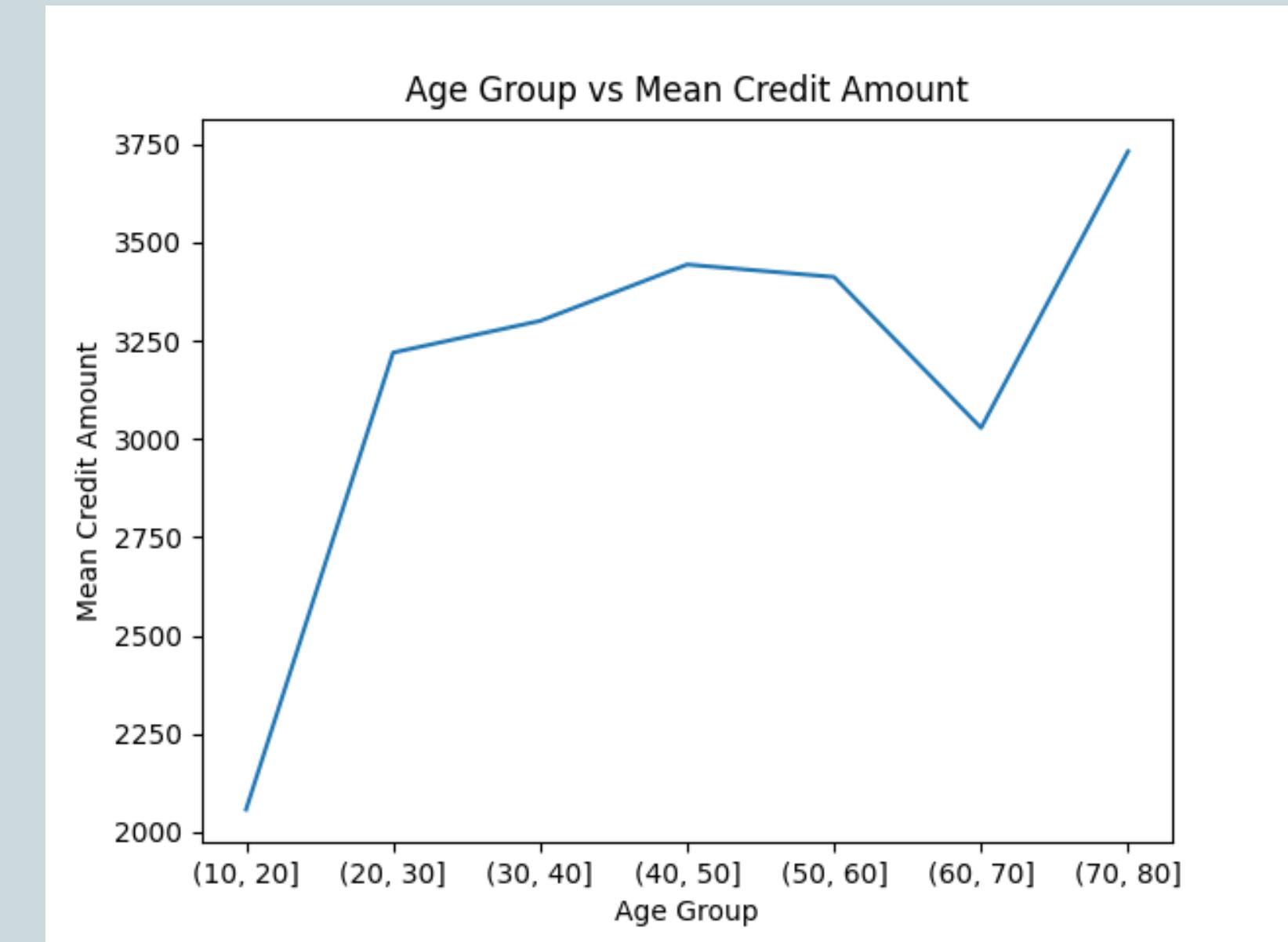
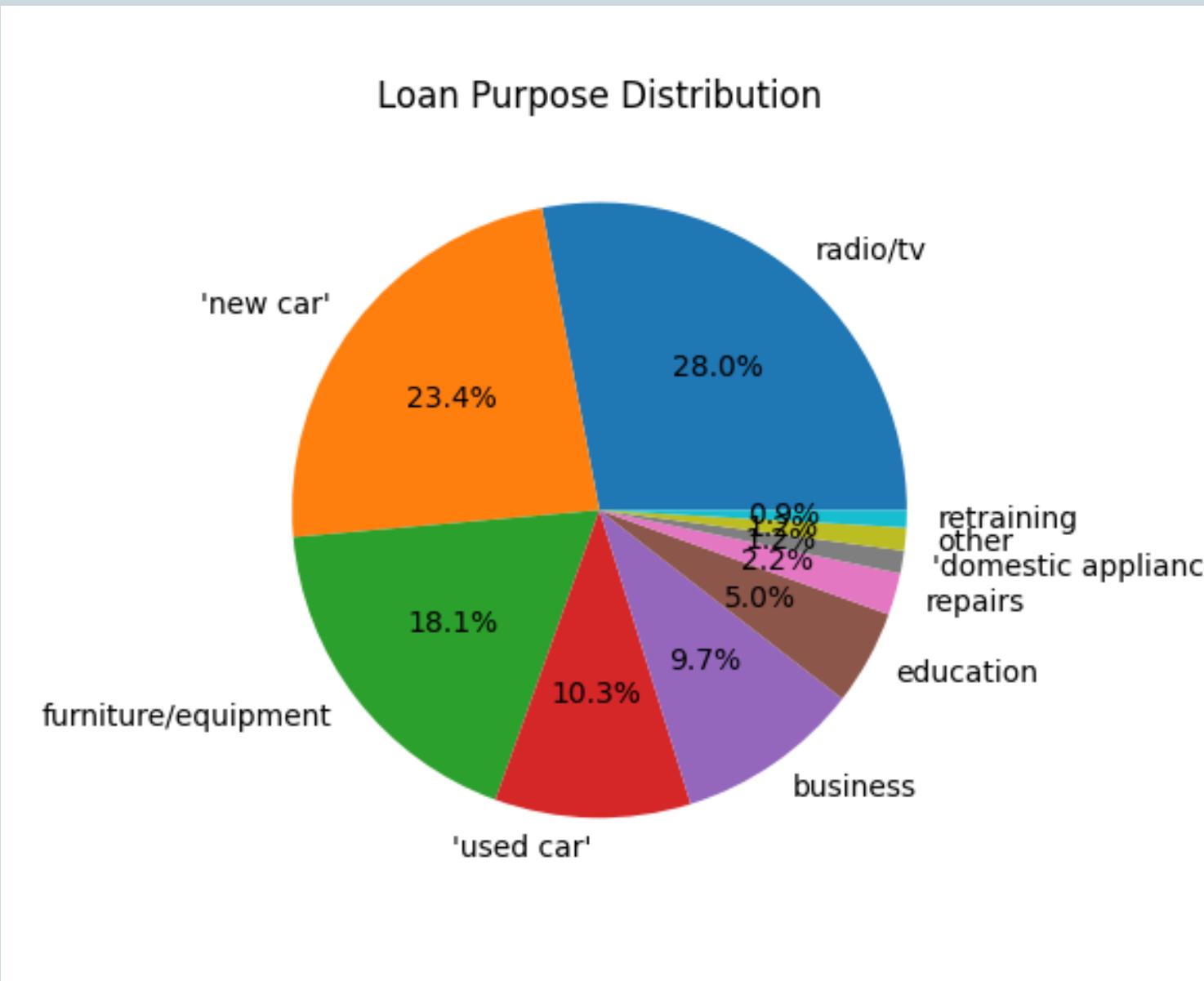
After

# 3. 探索資料

透過統計分析、資料視覺化，對客戶資料進行觀察

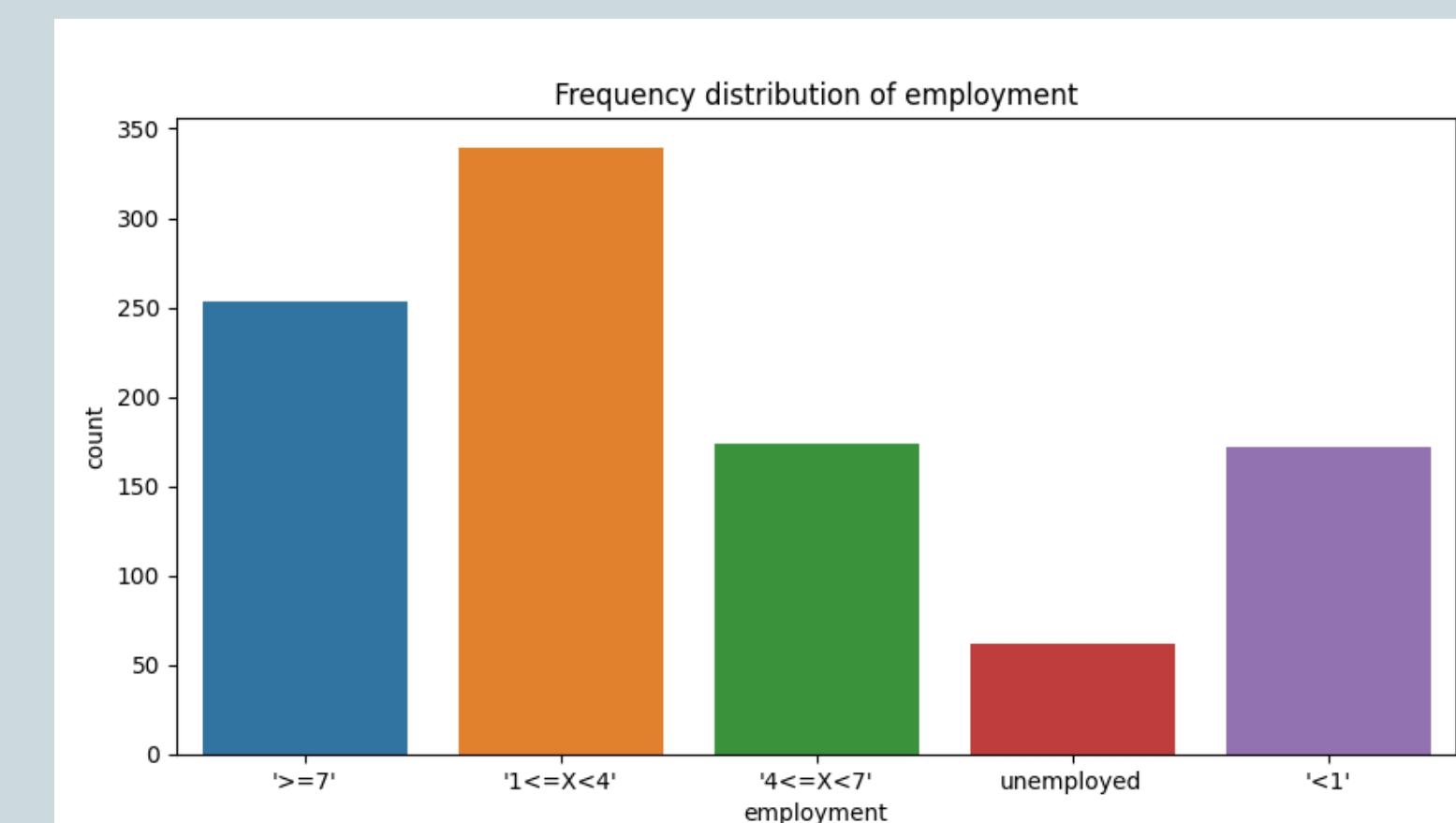
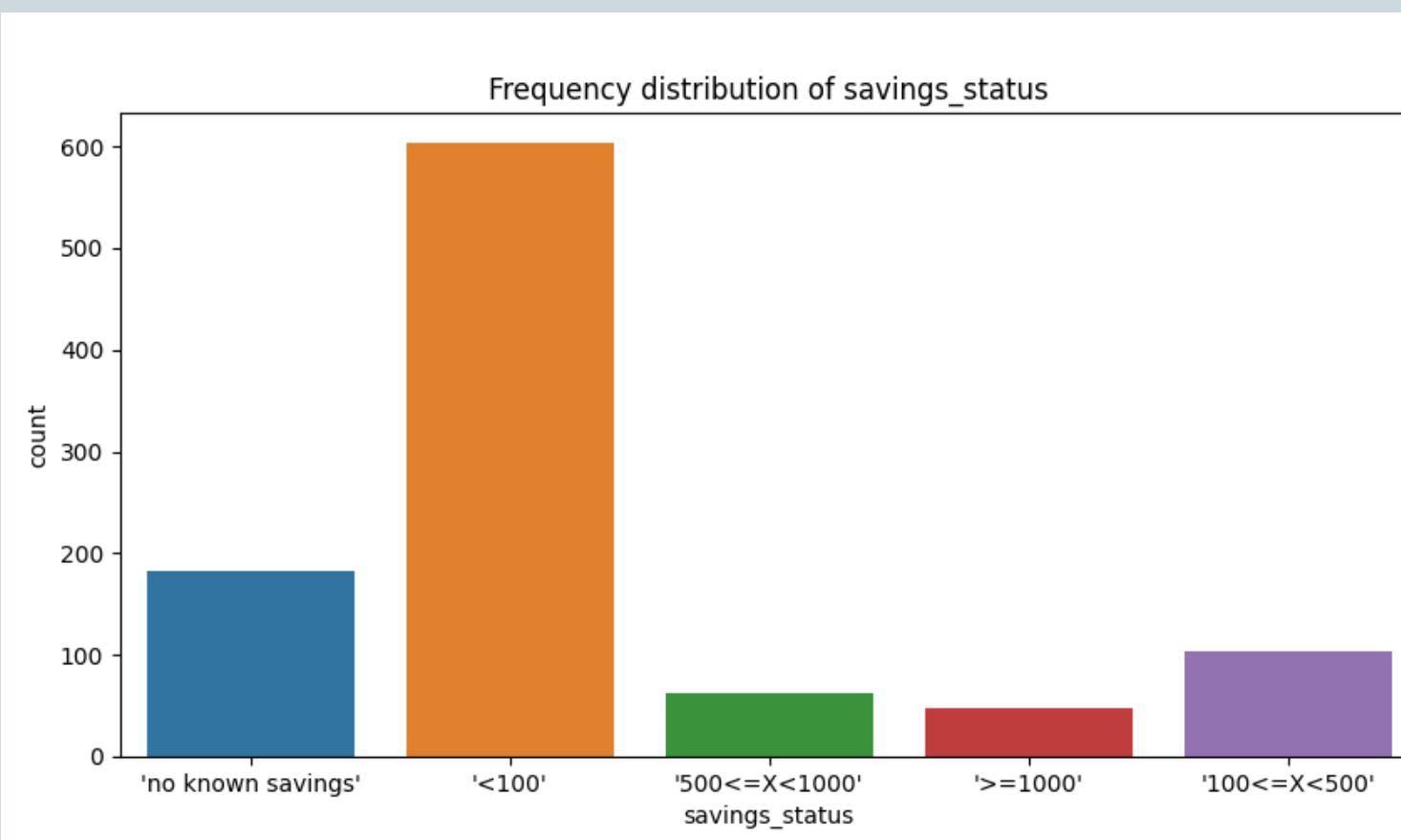
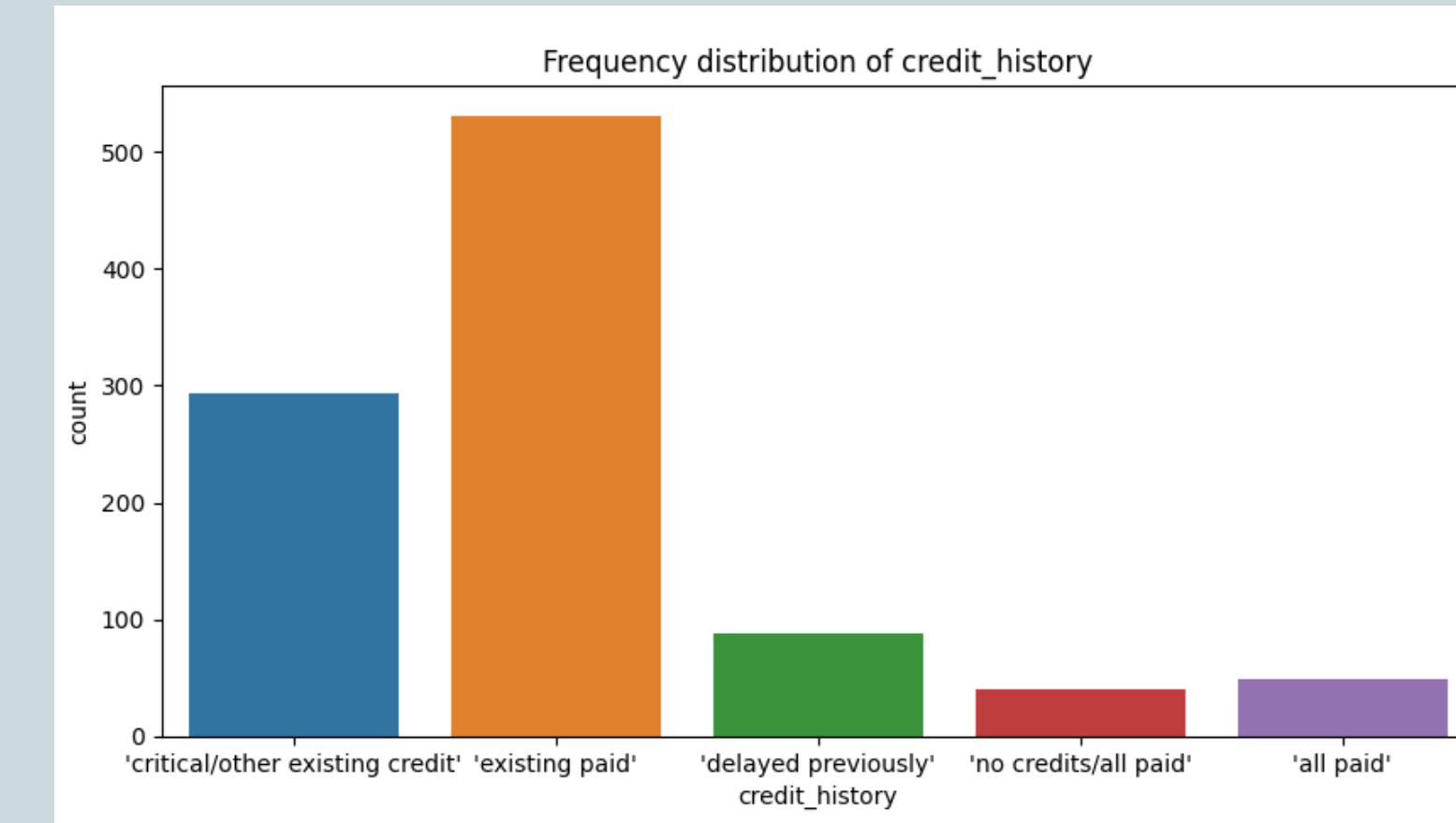
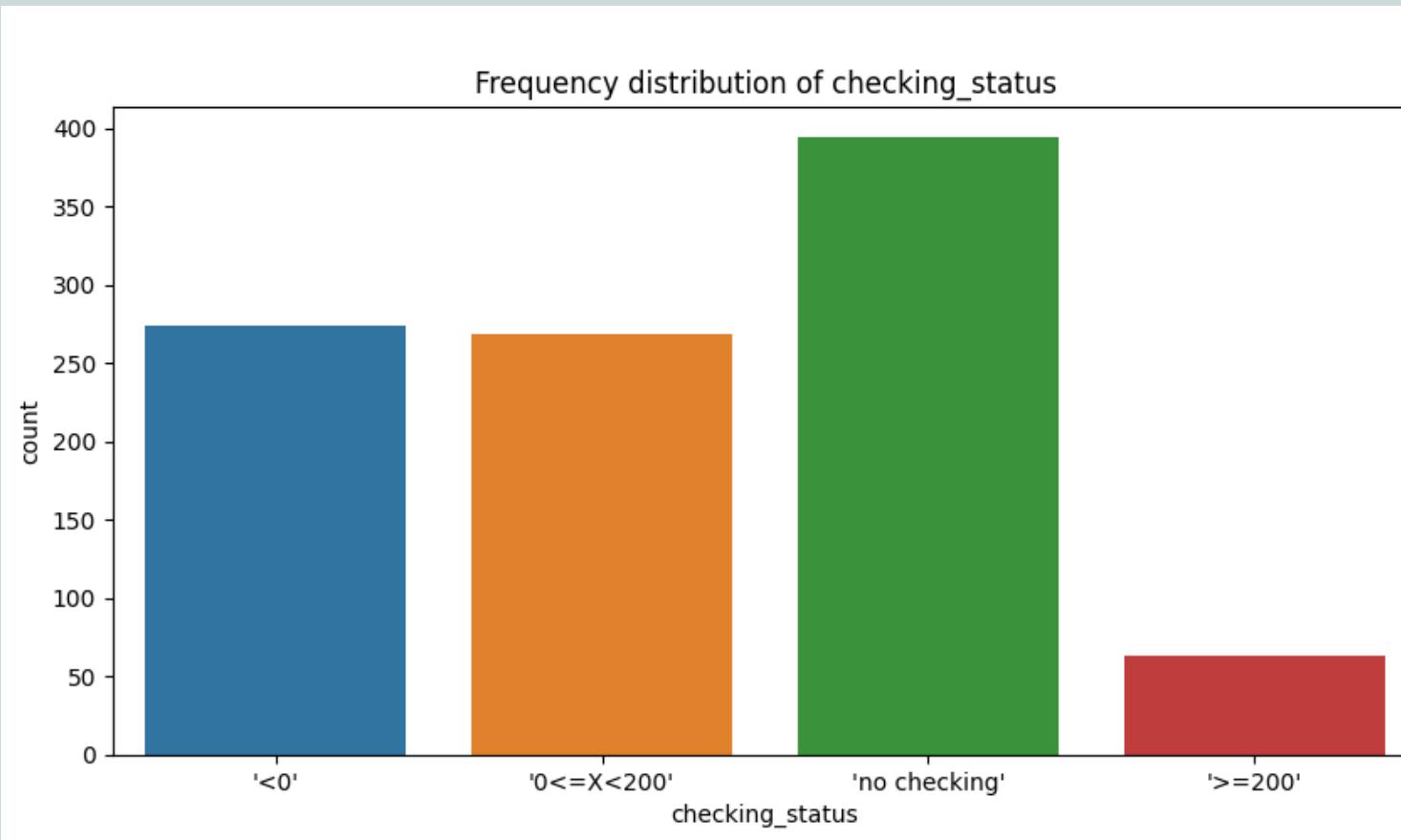


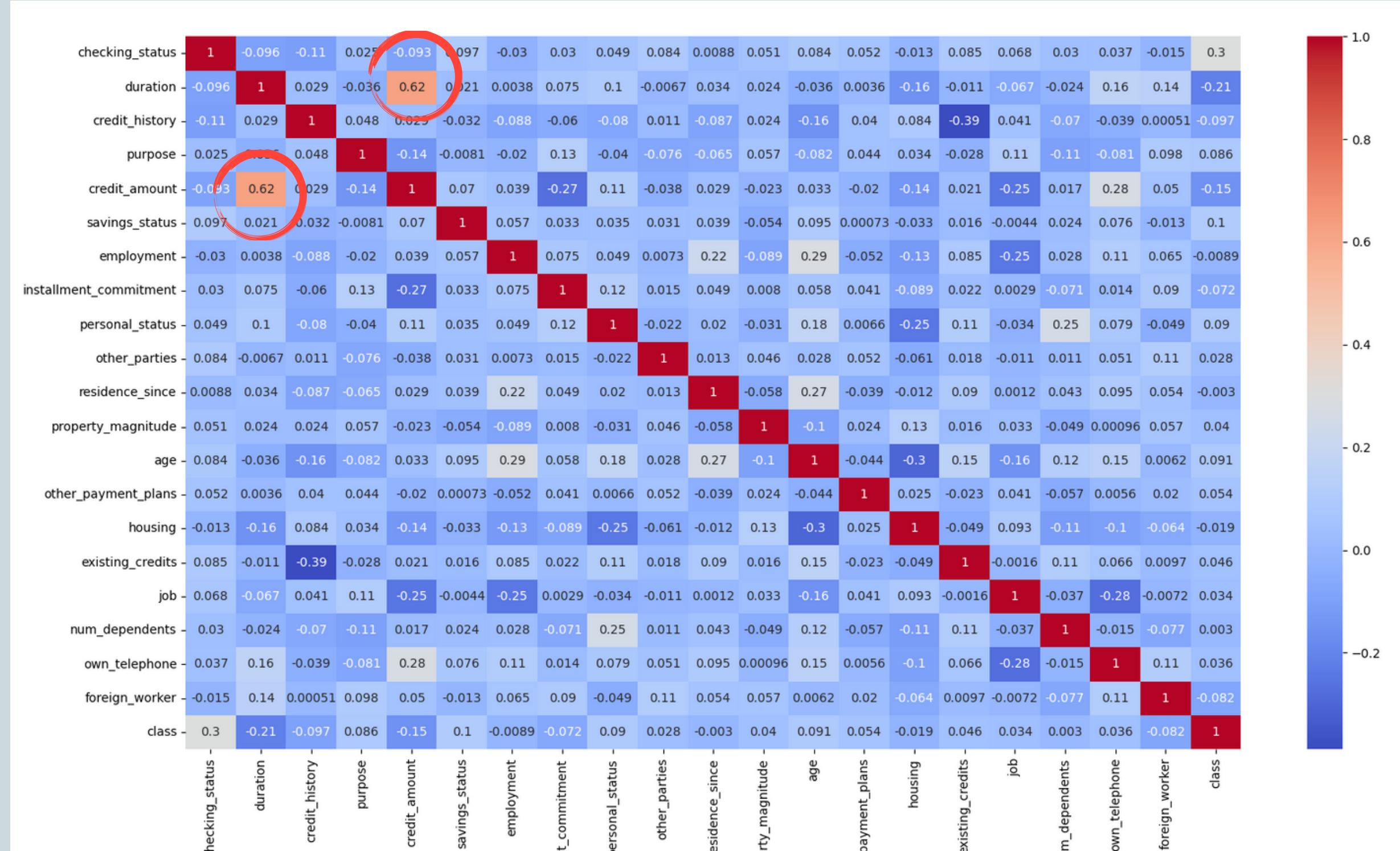
從兩張圖長條圖，我們可以發現貸款人大部分年齡都  
介於20到40歲，金額集中在3750德國馬克以下。



從圓餅圖中我們也可以看出，貸款目的主要為買車和電視還有家具。

從上面的圖中，可以觀察到隨著年齡增加，貸款金額也跟著成長，中位數大多落在於3250德國馬克附近。





貸款金額和償還時間的相關性特別高，可能是因為借款人需要在可接受的時間內還清借款，因此會選擇在可承受的範圍內選擇貸款金額。

# 4. 建立模型

選擇合適的分析方法和模型，建立預測模型或統計模型，以解釋資料和預測未來。

測試五個不同模型包括隨機森林、XGBoost、梯度提升樹、決策樹模型、邏輯回歸模型。選擇準確度和F1值兩個參數進行模型效能評估，發現隨機森林和XGBoost這兩個模型表現最好，選擇兩者進行效能測試。

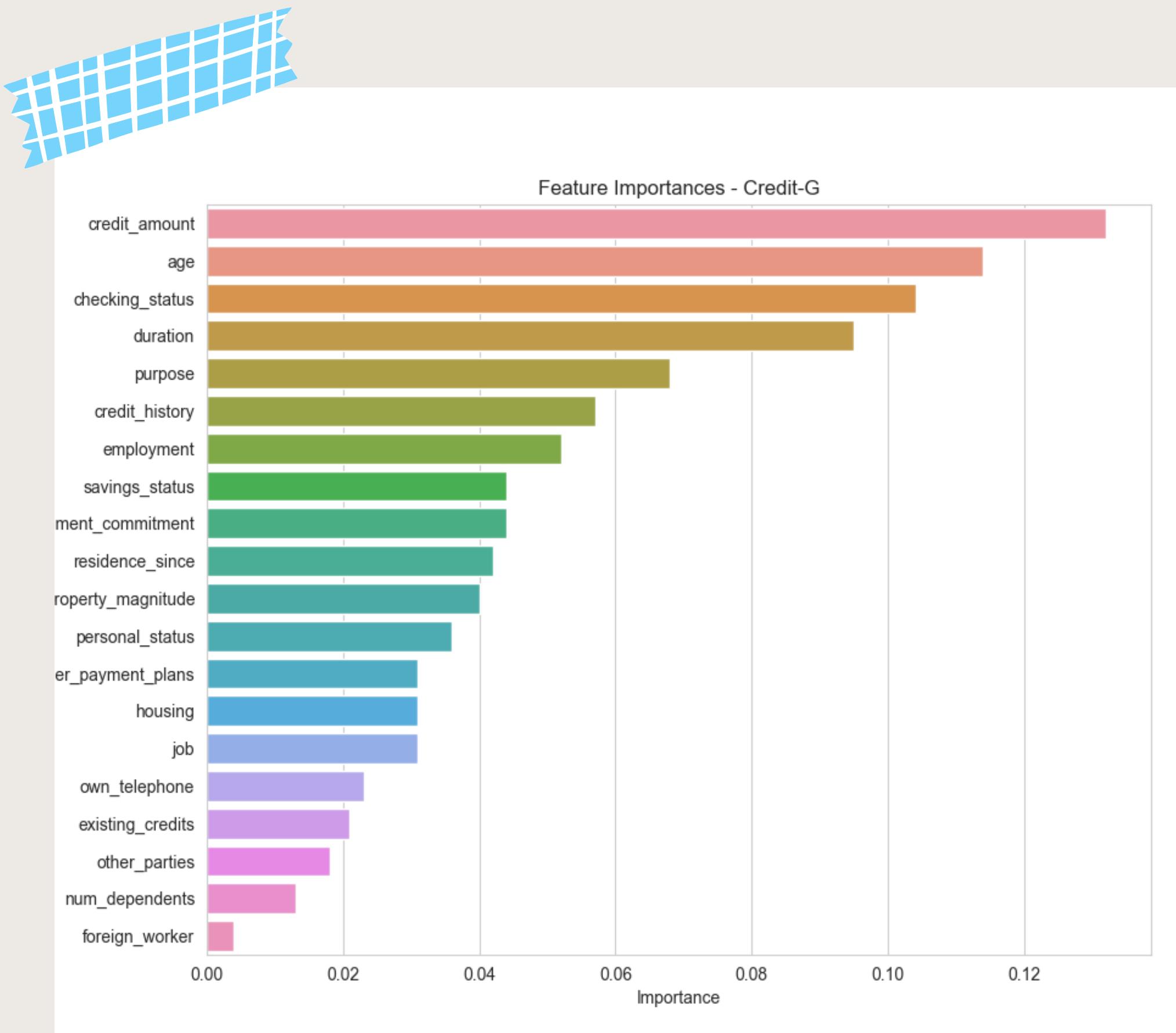
	準確度	F1值
隨機森林	0.79	0.86
XGBoost	0.8	0.86
梯度提升樹	0.78	0.86
決策樹模型	0.67	0.77
邏輯回歸模型	0.72	0.82

準確度：所有樣本的正確預測比例，即  $(TP + TN) / (TP + TN + FP + FN)$ 。

F1值：精確度和召回率的調和平均數，即  $2 * (precision * recall) / (precision + recall)$ 。

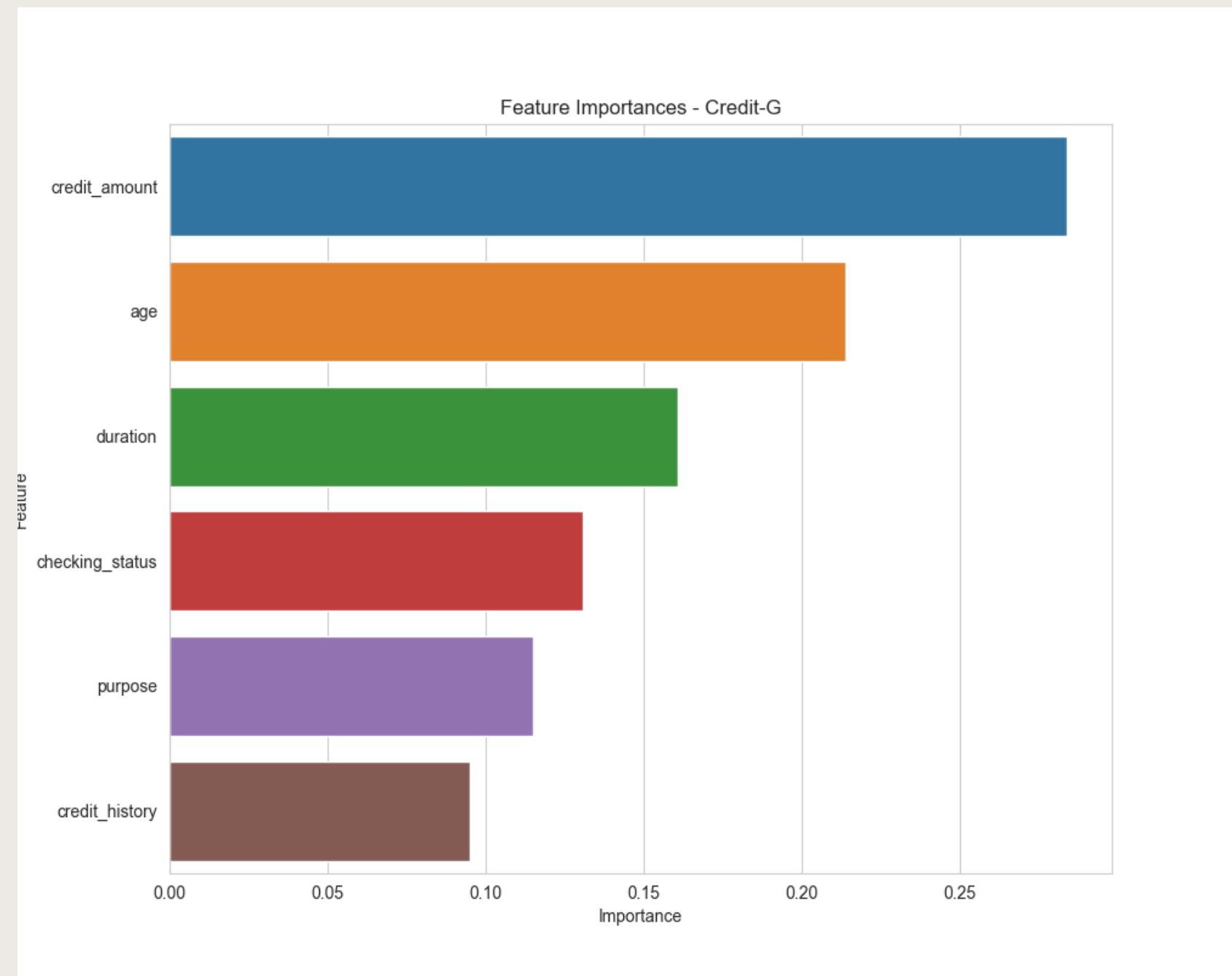
# 隨機森林模型

右邊這張圖是數據集中20個特徵的importance



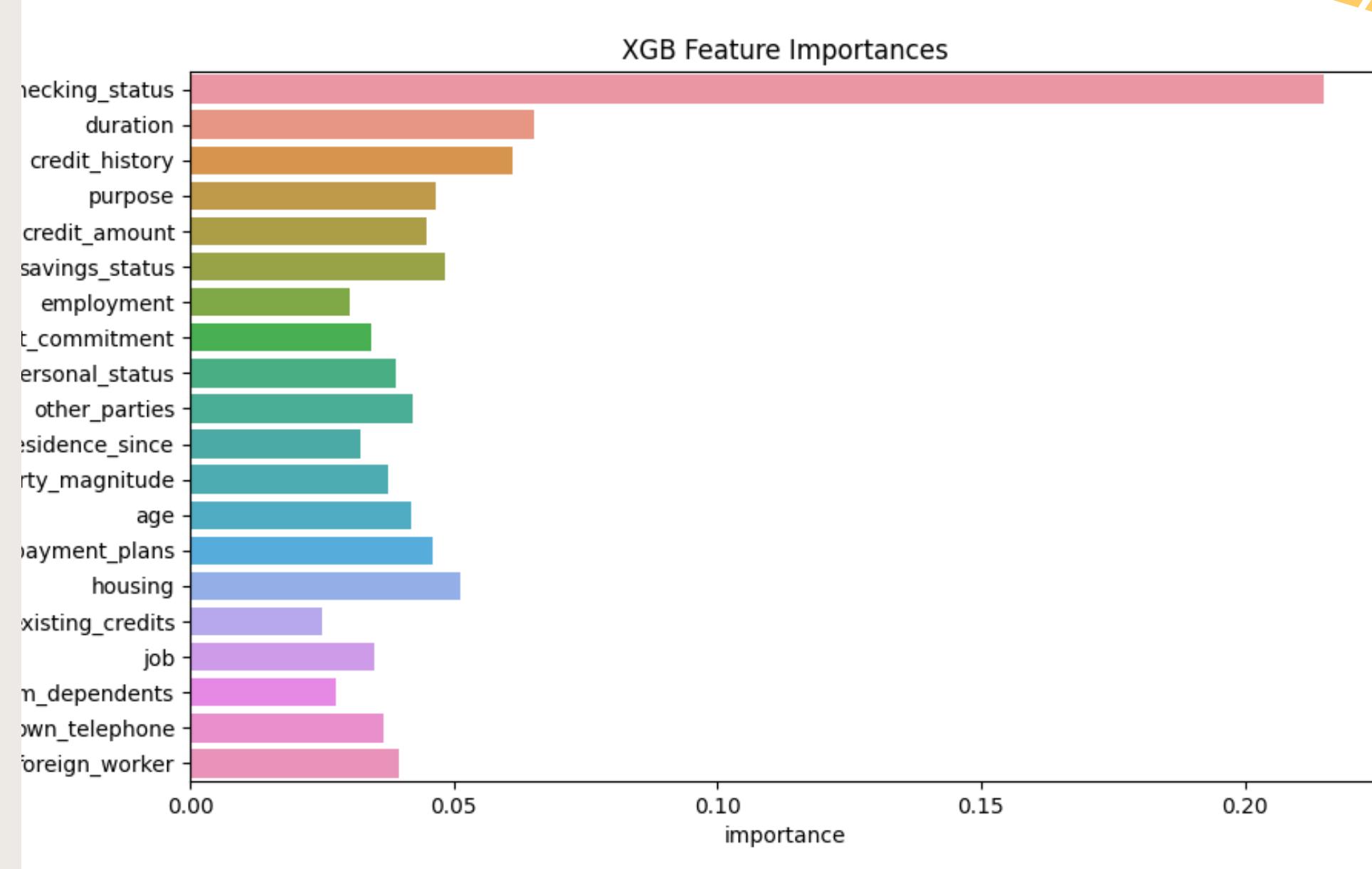
為了更近一步分析，選擇importance大於0.05的特徵再跑一次模型，發現準確率卻相同都是0.78。

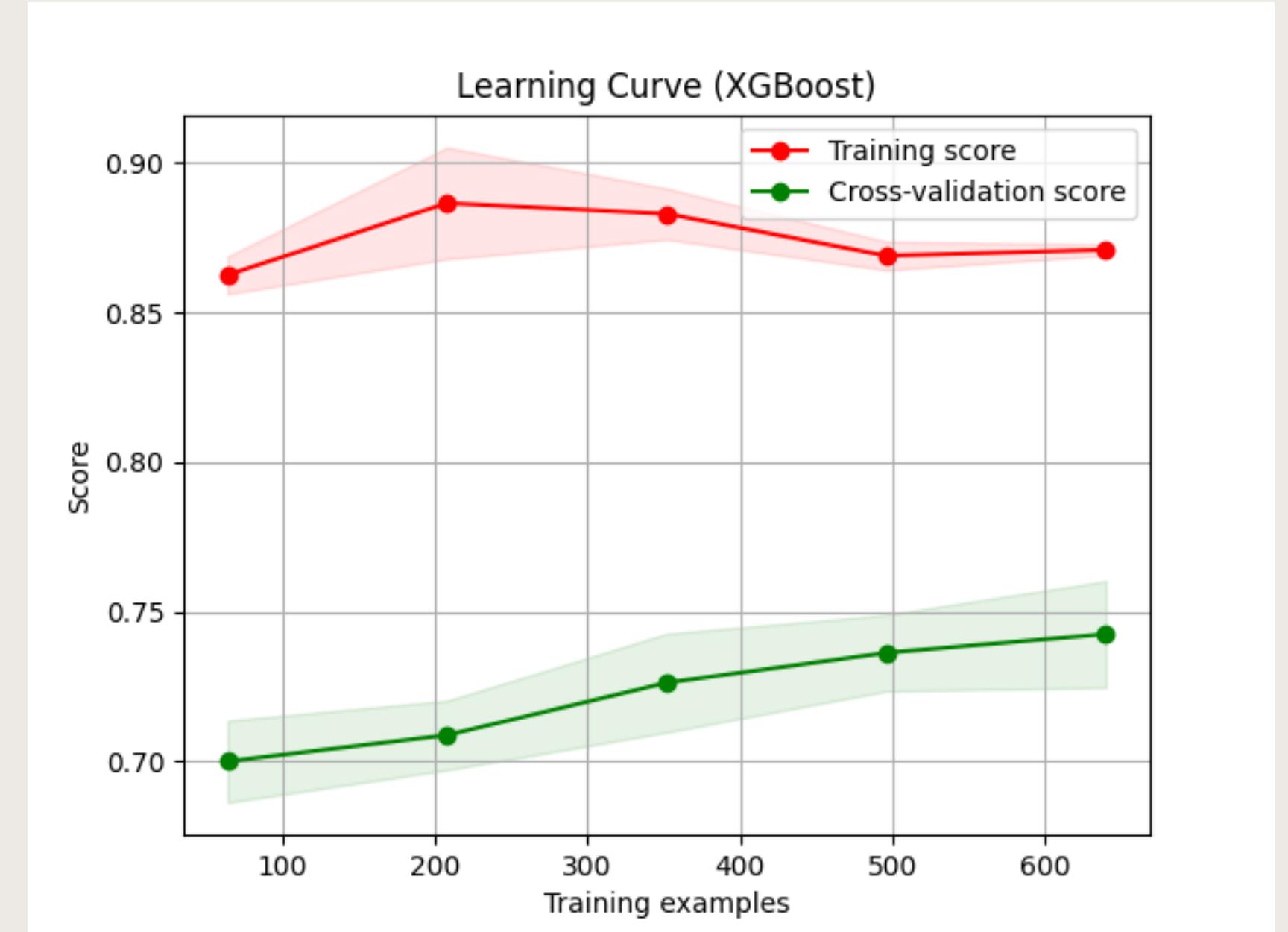
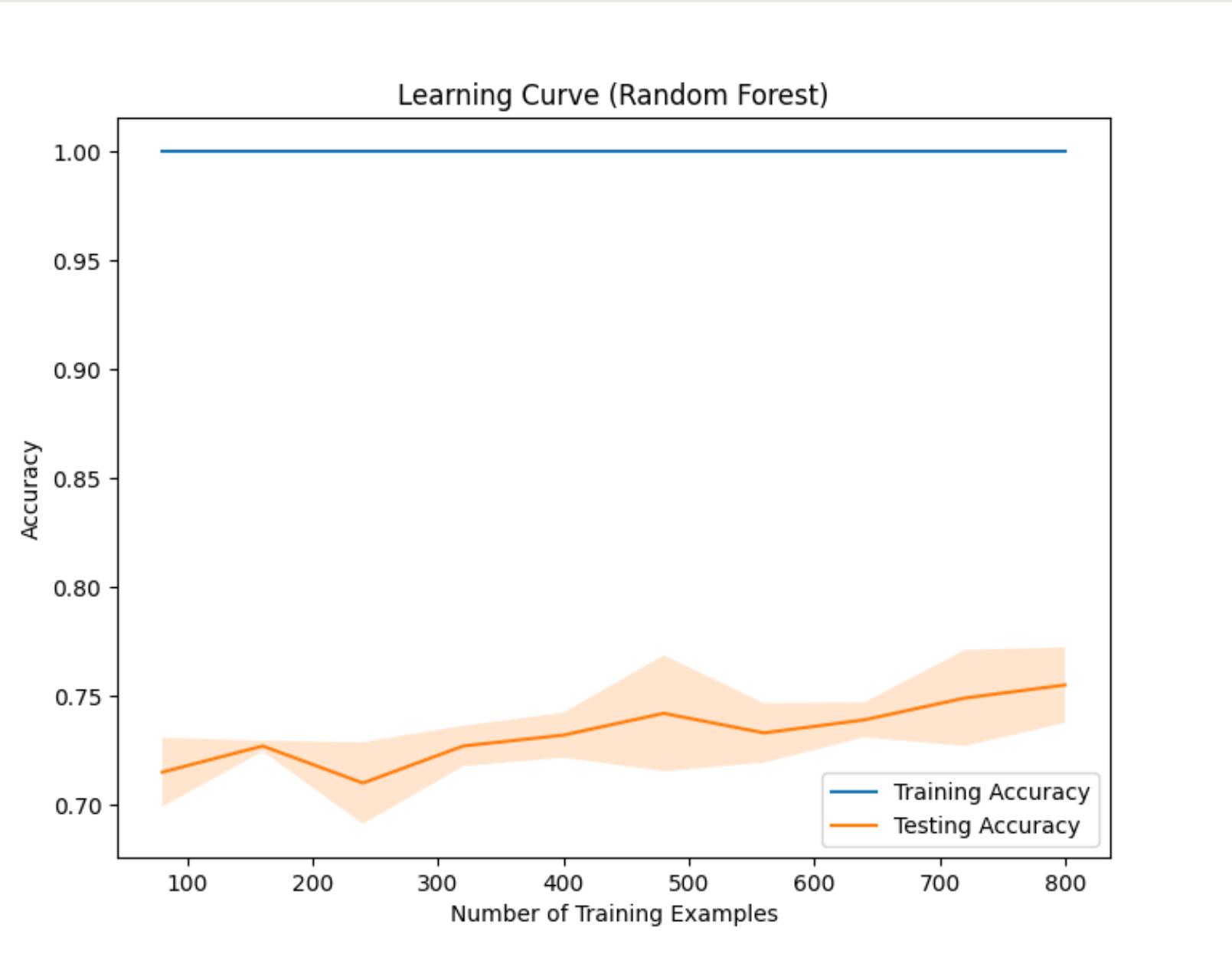
代表前六個特徵已經包含了足夠的信息，因此增加更多特徵並不能進一步提高模型性能。數據集中的噪聲或冗餘特徵對模型性能的影響較小，因此即使刪除部分特徵，模型仍然可以獲得類似的性能。



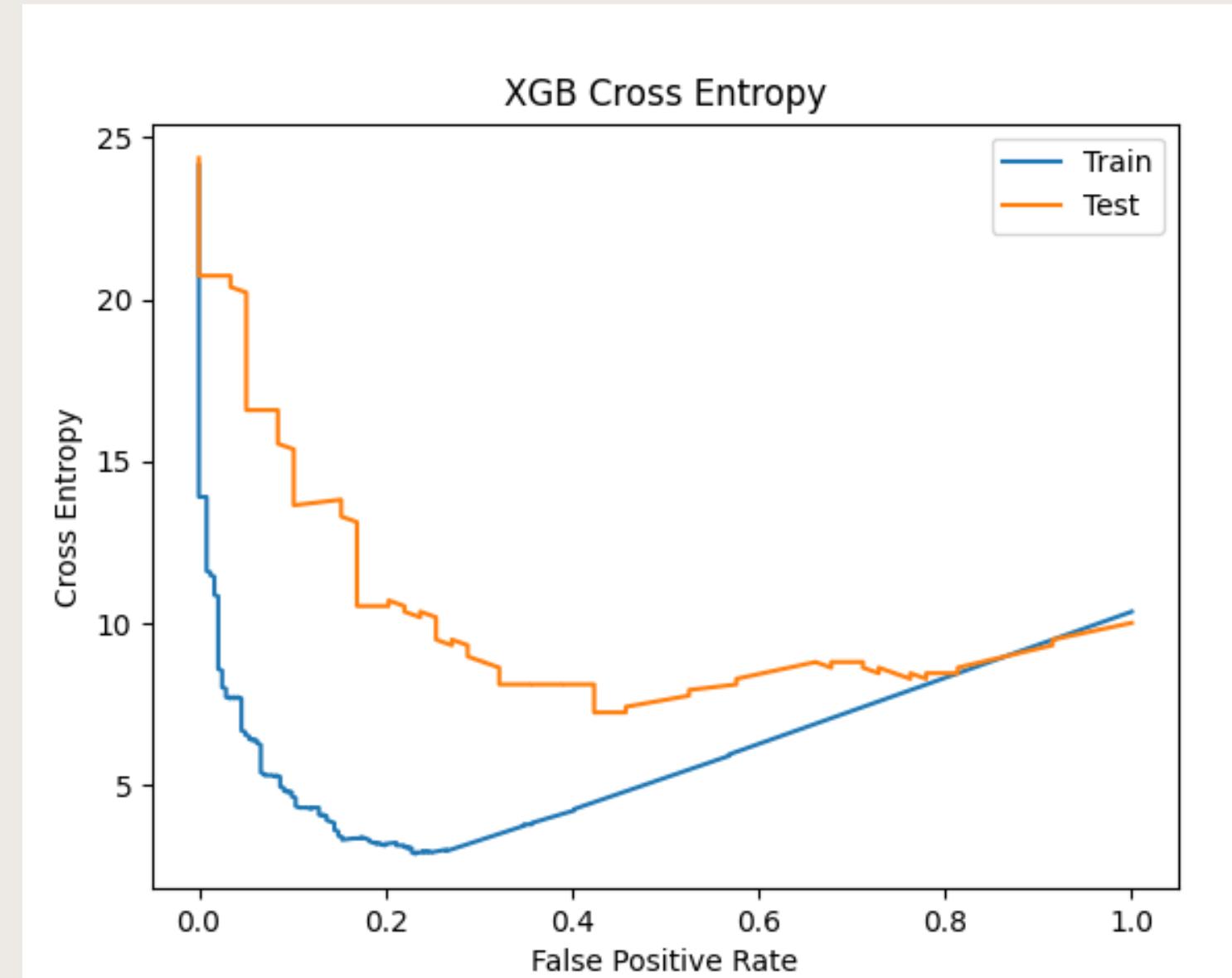
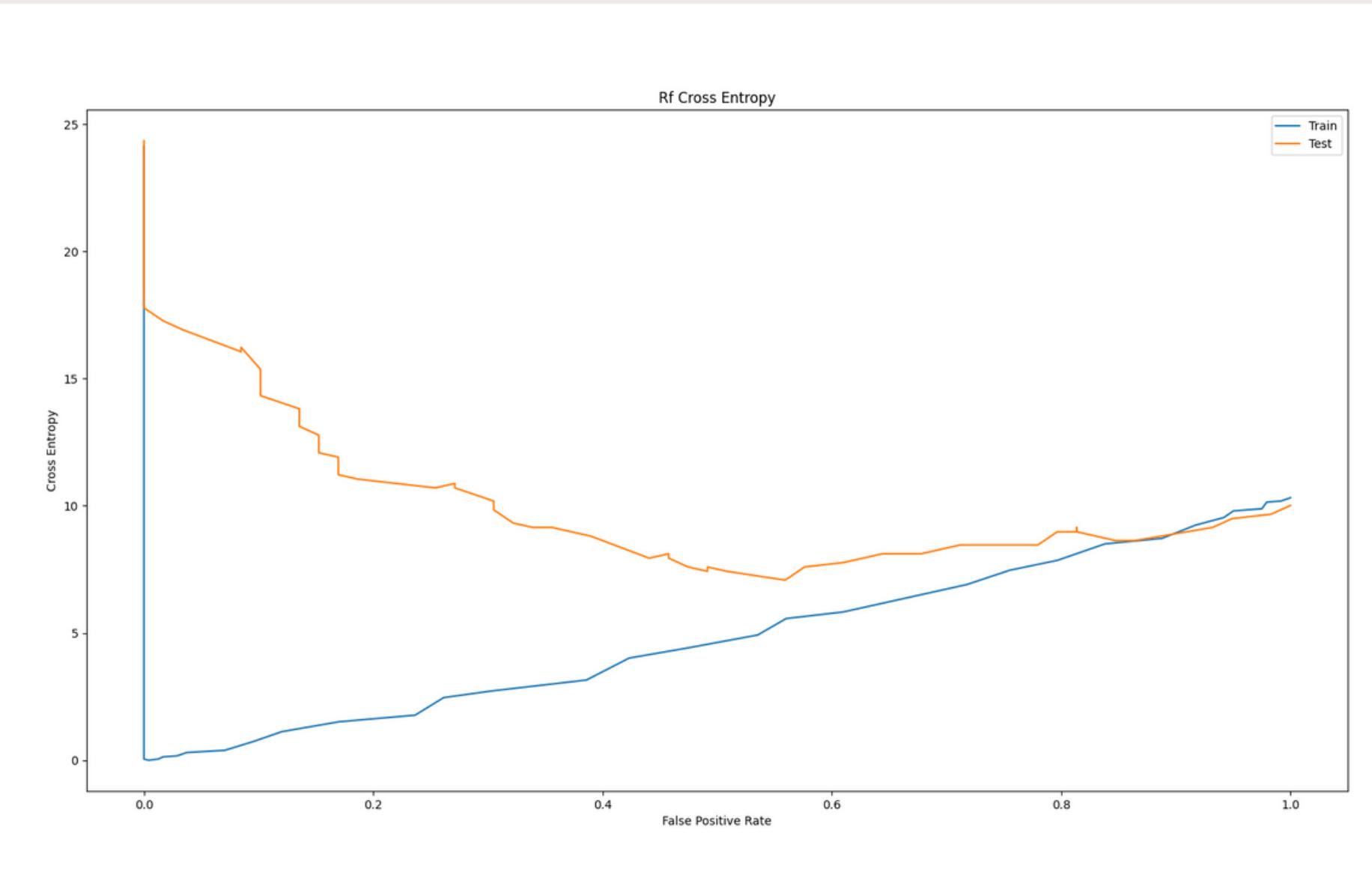
# XGB模型

同樣的方式也使用在XGB模型，也得出相同的結論。





從兩個模型的學習曲線，可以發現隨著樣本數增加，testing accuracy也持續增加，代表模型並沒有過度擬合。我也發現透過調整參數，準確率也只能維持在0.78和0.8之間，可能原因為數據量太少，需要加入新資料訓練才能增加模型效能。

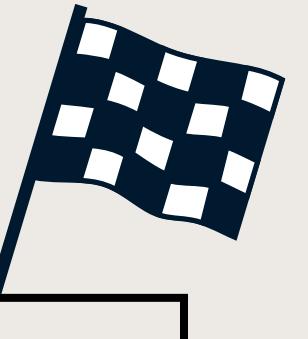


可以觀察到兩個模型在訓練集和測試集上的Cross-Entropy曲線，兩條線都逐漸重合，代表兩個模型都沒有over-fitting的問題，都具備很好的預測能力。

# 模型效能比較

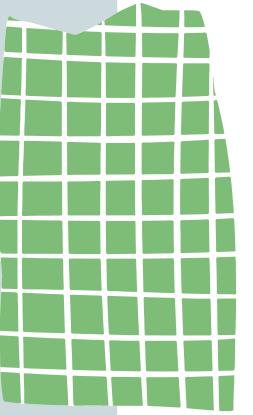
可以從各項評估指標，發現XGBoost模型比較較佳

	隨機森林	XGBoost
準確度	0.78	0.8 
F1 score	0.85	0.87 
Cross Entropy	0.48 	0.9
Cost loss	88	88



# 5. 檢驗結果

根據分析結果，選取在兩個模型都排名在前六名的特徵，分別為信用額度、貸款期限、信用記錄、貸款目的。



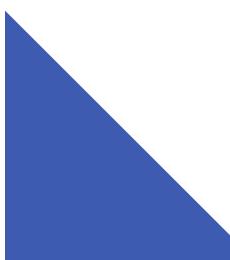
- 信用額度：代表著借款人的貸款金額，也代表著借款人的經濟實力，貸款金額較大的借款人可能在還款方面更有壓力。
- 貸款期限：貸款期限較長的借款人可能在還款方面存在風險，對於預測是否違約也有較大的影響力。
- 信用記錄：反映了一個人的借貸歷史，包括還款記錄、逾期記錄等信息，信用評分越高，個人的信用風險越低，越容易獲得貸款。
- 貸款目的：不同的貸款目的可能代表著不同的風險，對於預測是否違約也有一定的影響力。

---



# **Thank You!**

Thank you for participating.



**2023/05/09**

---