機器學習期末報告

一、 目的與動機

此資料集包含銀行客戶的個人資料與使用信用卡的相關資訊以及是否為流失客戶；此資料集可以用來預測客戶是否流失，這有助於銀行提前採取措施保留客戶。

二、 資料及介紹

1. 資料集名稱：Predicting Credit Card Customer Segmentation
2. 資料來源：Kaggle
3. 10,127 筆資料，20 個 features，1 個 label，沒有 missing value
4. X (Features) :

	欄位名稱	說明	資料型態
1	CLIENTNUM	客戶編號	Integer
2	Customer_Age	客戶年齡	Integer
3	Gender	性別	String
4	Dependent_count	受扶養人數	Integer
5	Education_Level	教育程度	String
6	Marital_Status	婚姻狀態	String
7	Income_Category	收入類別	String

8	Card_Category	卡片類別	String
9	Months_on_book	上書月數	Integer
10	Total_Relationship_Count	總關係數	Integer
11	Months_Inactive_12_mon	12 個月內閒置月數	Integer
12	Contacts_Count_12_mon	12 個月內聯絡次數	Integer
13	Credit_Limit	信用額度	Integer
14	Total_Revolving_Bal	總循環餘額	Integer
15	Avg_Open_To_Buy	平均開放購買比率	Integer
16	Total_Amt_Chng_Q4_Q1	第四季到第一季的總金額變動	Integer
17	Total_Trans_Amt	總交易金額	Integer
18	Total_Trans_Ct	總交易次數	Integer
19	Total_Ct_Chng_Q4_Q1	第四季到第一季的總次數變動	Integer
20	Avg_Utilization_Ratio	平均利用率比率	Integer

5. Y (Label) :

Attrition_Flag *(String)*：客戶是否流失

(2 *類*) Existing Customer (84%), Attrited Customer (16%)

三、 資料前處理

1. 將 CLIENTNUM 欄位刪除，因爲此欄位為客戶 ID。參與訓練的特徵為 19 個。

2. 確認資料集內沒有資料缺失

```
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

3. 將類別轉換為數值

將資料集的特定欄位使用 Label Encoder 轉換為數值表示。以下是轉換的欄位：

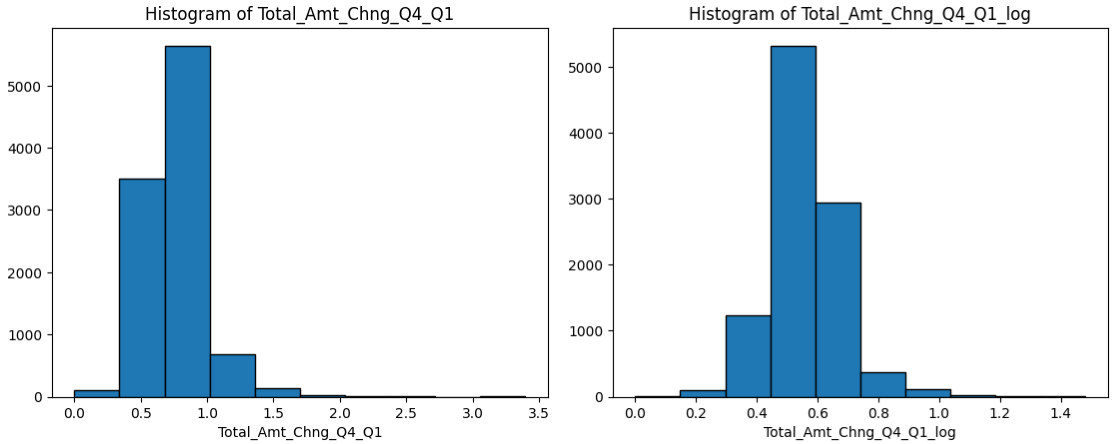
Education_Level, Marital_Status, Card_Category, Gender。

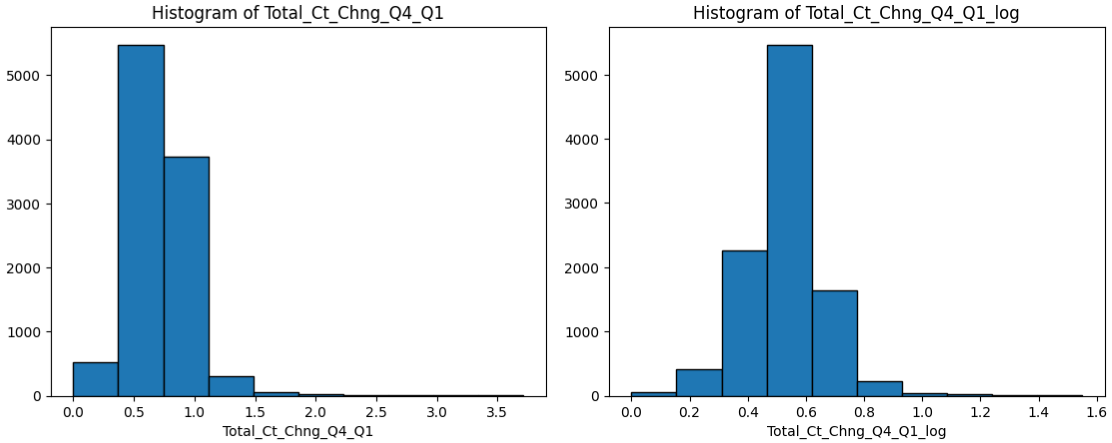
另外，針對 Income_Category 欄位，該欄位代表收入範圍，因此按照順序進行轉換。在此資料集中 Unknown 有 1,112 筆，故當成一個類別，收入範圍的順序為：
Unknown, Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +。最後，將 Attrition_Flag 欄位改 Existing Customer 為 0，Attrited Customer 為 1。

4. 檢查離群值

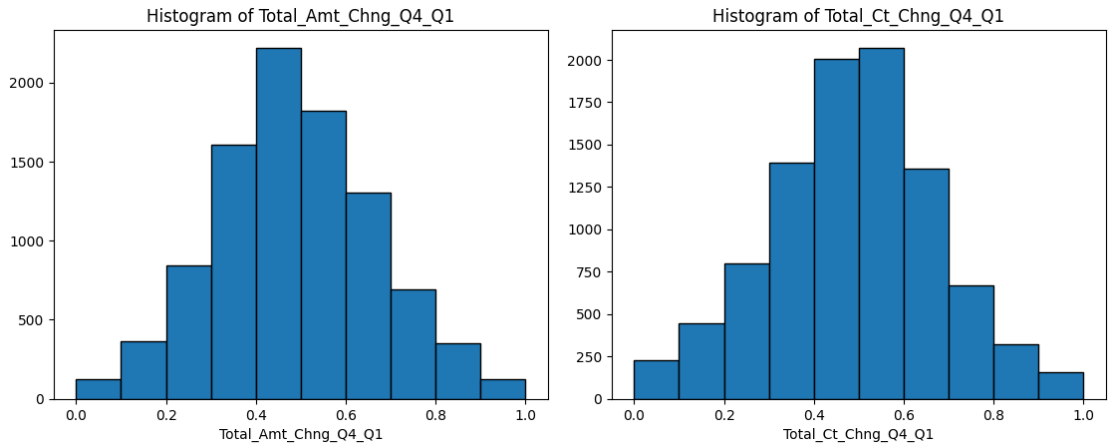
- 方法一取 log

觀察 Kaggle 的數據分佈圖，Total_Amt_Chng_Q4_Q1 與 Total_Ct_Chng_Q4_Q1 的資料分布偏左，因此取 log，左邊為處理前的資料分布圖，右邊為處理過的：



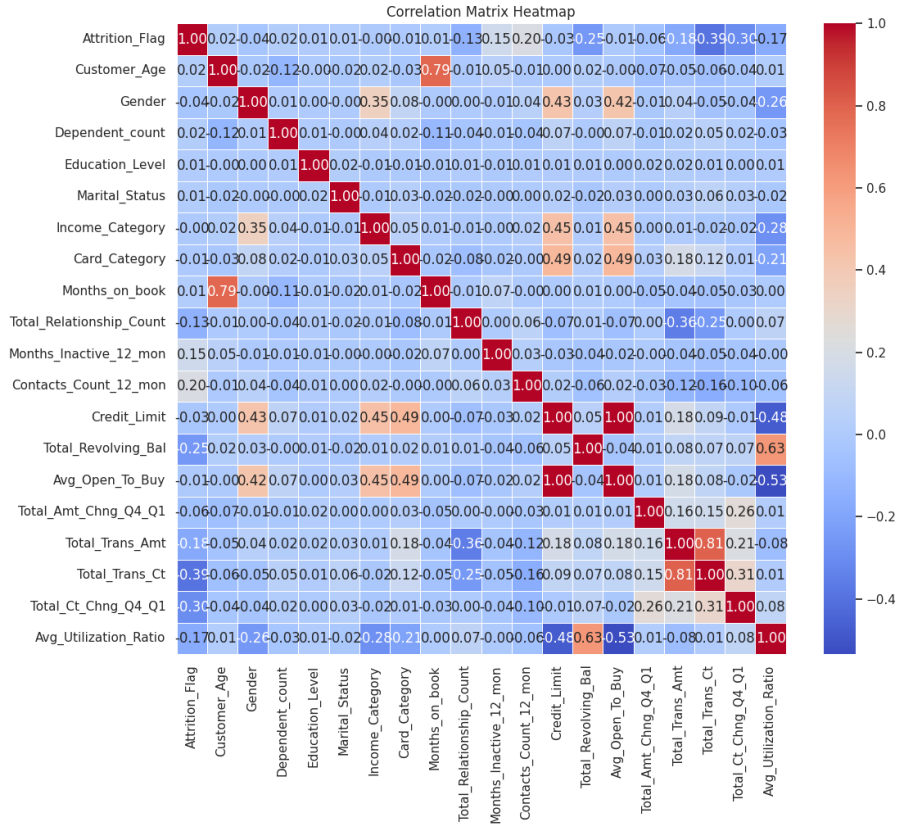


- 方法二四分位距



5. 特徵選擇

- 熱圖



可以看到 Credit_Limit(目前信用卡額度)與 Avg_Open_To_Buy(平均信用卡額度)相關程度是 1，非常合理，因此後面會刪除 Avg_Open_To_Buy

6. 資料正規化

使用 Max-Min 的方式將所有欄位的數值改為 0 到 1 之間。

四、 模型訓練與分析

首先，我們將資料分為訓練集和測試集，80%的資料用於模型的訓練，20%則用於測試。將 random_state 設置為 42，確保每次運行使用相同的隨機方式。下圖顯示，資料切割後，訓練集和測試集中的標籤比例相當。

```
Training Set Label Distribution:
0    0.839526
1    0.160474
Name: Attrition_Flag, dtype: float64

Testing Set Label Distribution:
0    0.838598
1    0.161402
Name: Attrition_Flag, dtype: float64
```

模型評估

- True Positive：模型預測為流失的客戶 (1)，且實際上是流失的。
- True Negative：模型預測為非流失的客戶 (0)，且實際上是非流失的。
- False Positive：模型預測為流失的客戶 (1)，但實際上是非流失的。
- False Negative：模型預測為非流失的客戶 (0)，但實際上是流失的。

研究目標是找出會流失的客戶。 希望**提高 True Positive，降低 False Negative**。因此主要以 **Recall** 及整體 **Accuracy** 的表現來評估模型。

1. Logistic Regression

Accuracy	Recall	Precision	F1 Score	TP	FN	TN	FP
0.8949	0.4862	0.7794	0.5989	159	168	1654	45

2. Bernoulli Naive Bayes

Accuracy	Recall	Precision	F1 Score	TP	FN	TN	FP
0.7828	0.5321	0.3774	0.4416	174	153	1412	287

3. Gaussian Naive Bayes

Accuracy	Recall	Precision	F1 Score	TP	FN	TN	FP
0.8776	0.6024	0.6254	0.6137	197	130	1581	118

4. SVM

	kernel	C	Accuracy	Recall	Precision	F1 Score	TP	FN

1	linear	1.0	0.8949	0.4893	0.7767	0.6004	160	167
2	rbf	1.0	0.8889	0.3639	0.8750	0.5140	119	208
3	poly	1.0	0.8766	0.3119	0.8031	0.4493	102	225
4	sigmoid	1.0	0.7485	0.2049	0.7458	0.2065	67	260
5	linear	5.0	0.8993	0.5199	0.7834	0.6250	170	157
6	rbf	5.0	0.9047	0.5107	0.8350	0.6338	167	160
7	poly	5.0	0.8934	0.4404	0.8136	0.5714	144	183
8	sigmoid	5.0	0.7428	0.2080	0.2601	0.2070	68	259
9	linear	10.0	0.8998	0.5229	0.7844	0.6275	171	156
10	rbf	10.0	0.9121	0.5627	0.8402	0.6740	184	143
11	poly	10.0	0.9003	0.4924	0.8173	0.6145	161	166
12	sigmoid	10.0	0.7399	0.2049	0.2006	0.2027	67	260
13	linear	100.0	0.9003	0.5260	0.7854	0.6300	172	155
14	rbf	100.0	0.9136	0.6024	0.8140	0.6924	197	130
15	poly	100.0	0.9164	0.6269	0.8008	0.7033	205	122
16	sigmoid	100.0	0.7399	0.2049	0.2006	0.2027	67	260

5. Voting classifier

- Logistic Regression, Bernoulli NB, Gaussian NB, SVM

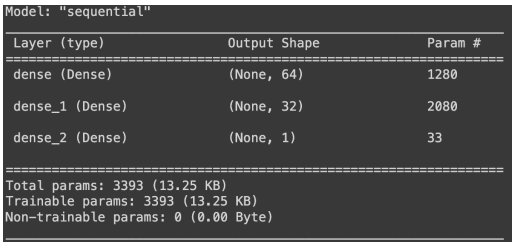
svm kernel		voting	weight	Accuracy	Recall	Precision	F1 Score	TP	FN
1	linear	hard	-	0.8929	0.4709	0.7778	0.5867	155	173
2	rbf	hard	-	0.8924	0.4128	0.8385	0.5533	135	192
3	poly	hard	-	0.8894	0.4067	0.8160	0.5429	133	194
4	sigmoid	hard	-	0.8806	0.4895	0.7891	0.3547	116	211
5	linear	soft	1, 1, 1, 1	0.8978	0.4985	0.7913	0.6116	163	164
6	rbf	soft	1, 1, 1, 1	0.8998	0.5046	0.8010	0.6191	165	162
7	poly	soft	1, 1, 1, 1	0.8978	0.5015	0.7885	0.6131	164	163
8	sigmoid	soft	1, 1, 1, 1	0.8880	0.3700	0.8521	0.5160	121	206
9	linear	soft	2, 1, 1, 1	0.8973	0.5046	0.7820	0.6134	165	162
10	rbf	soft	2, 1, 1, 1	0.8988	0.5015	0.7961	0.6154	164	163
11	poly	soft	2, 1, 1, 1	0.8973	0.5015	0.7847	0.6119	164	163
12	linear	soft	1, 2, 1, 1	0.8978	0.4801	0.8093	0.6027	157	170
13	rbf	soft	1, 2, 1, 1	0.8944	0.4618	0.7989	0.5853	151	176
14	poly	soft	1, 2, 1, 1	0.8954	0.4679	0.8010	0.5907	153	174
15	linear	soft	1, 1, 2, 1	0.8954	0.5229	0.7533	0.6173	171	156
16	rbf	soft	1, 1, 2, 1	0.8944	0.5107	0.7557	0.6095	167	160
17	poly	soft	1, 1, 2, 1	0.8954	0.5199	0.7556	0.6159	170	157

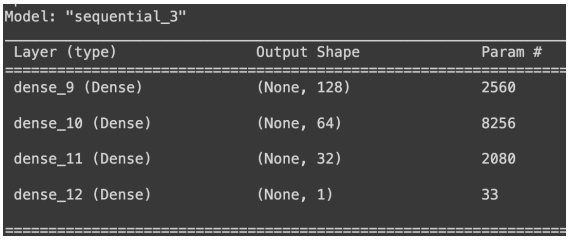
18	linear	soft	1, 1, 1, 2	0.8983	0.5076	0.7867	0.6171	166	161
19	rbf	soft	1, 1, 1, 2	0.8993	0.5046	0.7971	0.6180	165	162
20	poly	soft	1, 1, 1,2	0.8954	0.4924	0.7778	0.6030	161	166

以上幾種模型在此資料集的 Accuracy 雖然偏高，但 Recall 最高僅 0.6269，表示模型在發現流失客戶方面表現不佳。分析表現最好的模型發現有 122 個實際會流失的客戶未被成功預測，表示以上模型不適合用來預測客戶是否會流失。

6. MLNN

使用 TensorFlow 框架建立兩個神經網絡模型，左邊的模型架構有三個全連接層，右邊則是有四個全連接層。





	架構	Epoch	Accuracy	Recall	Precision	F1 Score	TP	FN
1	左	100	0.9373	0.7676	0.8311	0.7981	251	76
2	左	300	0.9339	0.7982	0.7933	0.7957	262	65
3	左	500	0.9373	0.8196	0.7976	0.8084	268	59
4	右	100	0.9265	0.8379	0.7405	0.7862	274	53
5	右	300	0.9388	0.7829	0.8285	0.8050	256	71
6	右	500	0.9413	0.7768	0.8467	0.8102	254	73

兩種架構建立的模型表現 Recall 皆有提升，最高有到 0.8266。

測試 Total_Amt_Chng_Q4_Q1 與 Total_Ct_Chng_Q4_Q1 沒有取 log 的預測結果如下

	架構	Epoch	Accuracy	Recall	Precision	F1 Score	TP	FN
1	左	100	0.9294	0.7431	0.8046	0.7727	243	84
2	左	300	0.9269	0.8043	0.7579	0.7804	263	64

3	**左**	**500**	0.9408	0.7645	0.8532	0.8065	250	77
4	**右**	**100**	0.9432	0.8287	0.8212	0.8250	271	56
5	**右**	**300**	0.9442	0.8073	0.8408	0.8237	264	63
6	**右**	**500**	0.9462	0.8104	0.8494	0.8294	265	62

五、 結論

	Accuracy	**Recall**	**TP**	**FN**
Logistic Regression	0.8949	0.4862	159	168
Bernoulli Naive Bayes	0.7828	0.5321	174	153
Gaussian Naive Bayes	0.8776	0.6024	197	130
SVM	0.9164	0.6269	205	122
Voting classifier	0.8954	0.5229	171	156
MLNN	0.9492	0.8226	269	58