

112-1

期末資料分析競賽

Final Data Analysis Competition

分析主題說明

Dataset :

- 顧客基本資料 : **customer_data.csv**
- 顧客郵遞區號 : **customer_zip.csv**
- 顧客資料紀錄 : **7,043 筆** , **40 欄位**

主題 (問題)說明

1. 探索性分析所有顧客資料的特徵(視覺化呈現)，並處理資料的格式與問題(例如: 遺失值等)。
2. 針對所有城市(City)，計算總收入前3高的城市，個別分析其總收入、總費用及總退款的金額等重要特徵。
3. 根據顧客的狀態(目標欄位)建立其相關的決策樹及規則，並分析及評估決策樹的效能等相關指標(例如:正確率等)。
4. 針對顧客所屬地區的經緯度，將顧客依地理位置分群，並比較不同群組特徵(例如：性別、年齡、婚姻、扶養人數等)的差異性，針對其中任一群組，再建立其群組使用公司服務的關聯規則。
5. 針對顧客的重要特徵分群，找出2~3群最有特色的顧客，並解釋其價值與意義。
6. 根據顧客年齡的差異(分成老、中、青)，比較其使用公司服務的關聯規則的異同。
7. 自由主題(自選問題 1~N；其中一題需與郵遞區號有關)。

評比規則

- **截止日期**：2023/12/17 (日) 23:30
- **繳交內容**：
 - **Codes**：python, 可以分成幾個檔案或單一檔案。
 - **Presentation**：PPT, (30mins + 10mins QA)
- **評比標準**：
 - **模型評估結果**：**50 pts**. Precision, F1, AUC, Confidence, ...
 - **口頭報告技巧**：**30 pts**
 - **分析流程 & 方法運用**：**10 pts**
 - **Codes品質**：**10 pts**

欄位說明 (1/4)

客戶編號	識別每個客戶的唯一 ID
性別	顧客性別：男、女
年齡	客戶在財政季度結束時的當前年齡（ 2022 年第 2 季 ）
婚姻	表示客戶是否已婚：是、否
扶養人數	表示與客戶同住的家屬人數（家屬可以是孩子、父母、祖父母等）
城市	客戶主要居住城市（加州）
郵遞區號	客戶主要居住地的郵遞區號
緯度	客戶主要居住地的緯度
經度	客戶主要居住地的經度
推薦次數	表示迄今為止客戶將朋友或家人推薦給該公司的次數
加入期間 (月)	表示截至上述指定季度末客戶在公司工作的總月數
優惠方式	最後一個識別客戶接受的行銷優惠：無、優惠 A 、優惠 B 、優惠 C 、優惠 D 、優惠 E
電話服務	指示客戶是否向公司訂購家庭電話服務：是、否
平均長途話費	表示客戶的平均長途話費，計算至上述指定季度末（如果客戶未訂閱家庭電話服務，則為 0 ）

欄位說明 (2/4)

多線路服務	表示客戶是否與公司訂閱了多條電話線路：是、否（如果客戶未訂閱家用電話服務，則為否）
網路服務	指示客戶是否向公司訂購 Internet 服務：是、否
網路連線類型	指示客戶的網路連線類型： DSL 、光纖、有線電視（如果客戶未訂閱網路服務，則為「無」）
平均下載量 (GB)	表示客戶的平均下載量（以 GB 為單位），計算到上述指定季度末（如果客戶未訂閱網路服務，則為 0 ）
線上安全服務	表示客戶是否訂閱了本公司提供的額外線上安全服務：是、否（如果客戶未訂閱網路服務，則為否）
線上備份服務	表示客戶是否訂閱了本公司提供的額外線上備份服務：是、否（如果客戶未訂閱網路服務，則為否）
設備保護計劃	表示客戶是否為其互聯網設備訂閱了公司提供的附加設備保護計劃：是、否（如果客戶未訂閱互聯網服務，則為否）
技術支援計劃	指示客戶是否訂閱了公司的附加技術支援計劃以減少等待時間：是、否（如果客戶未訂閱網路服務，則為否）

欄位說明 (3/4)

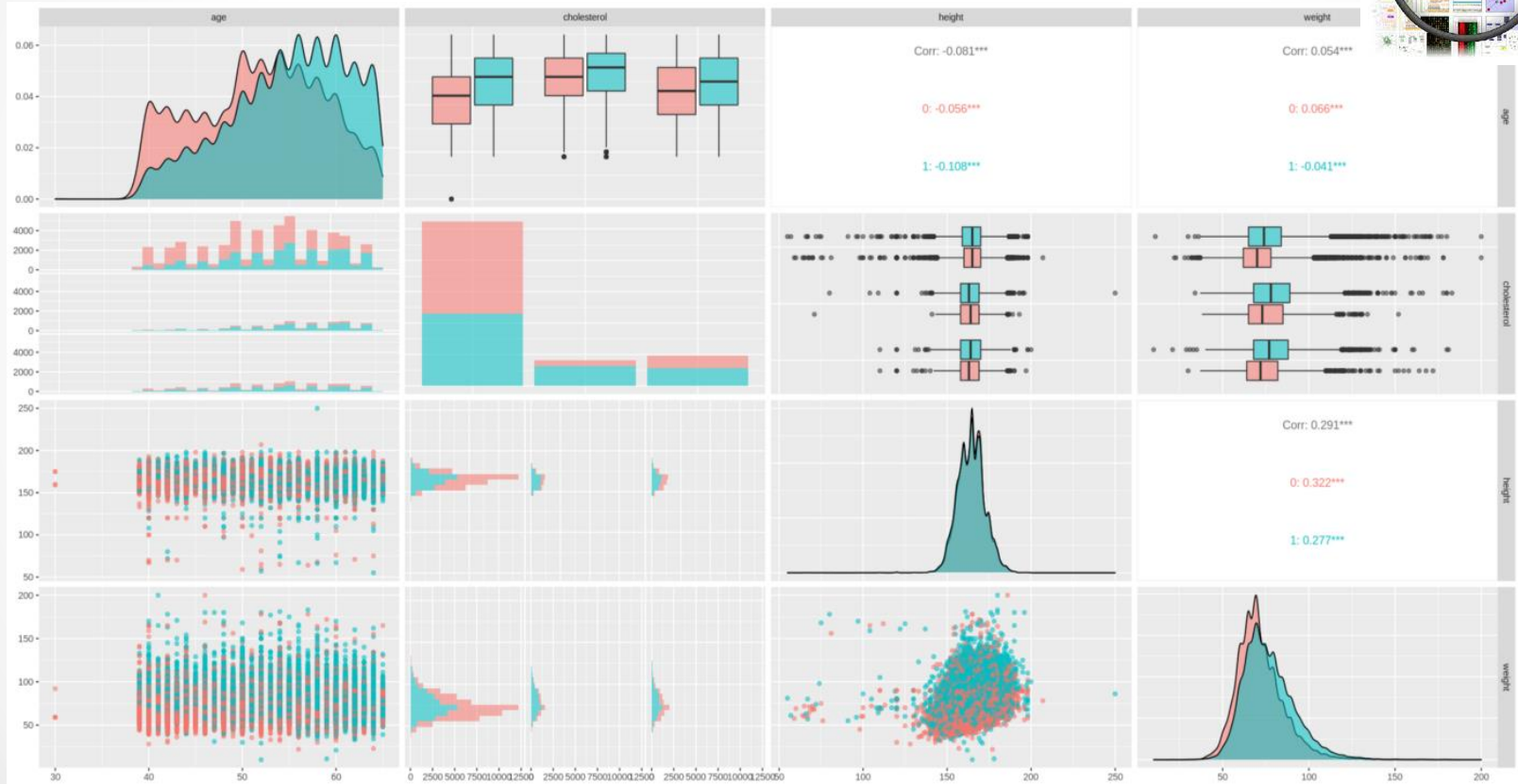
電視節目	指示客戶是否使用其網路服務免費觀看第三方供應商的電視節目：是、否（如果客戶未訂閱網路服務，則為否）
電影節目	指示客戶是否使用其網路服務免費觀看第三方供應商的電影：是、否（如果客戶未訂閱網路服務，則為否）
音樂節目	指示客戶是否使用其網路服務免費播放第三方供應商的音樂：是、否（如果客戶未訂閱網路服務，則為否）
無限資料下載	表示客戶是否已支付額外的月費以獲得無限制的資料下載/上傳：是、否（如果客戶未訂閱網路服務，則為否）
合約類型	表示客戶目前的合約類型：按月、一年、兩年
無紙化計費	指示客戶是否選擇無紙化計費：是、否
支付帳單方式	指示客戶如何支付帳單：銀行提款、信用卡、郵寄支票
每月費用	表示客戶目前每月從本公司獲得的所有服務的總費用
總費用	表示客戶的總費用，計算至上述指定季末
總退款	表示客戶的退款總額，計算至上述指定季末

欄位說明 (4/4)


額外數據費用	表示截至上述指定季末客戶超出其計劃中指定的額外資料下載的總費用
額外長途費用	表示截至上述指定季度末，客戶的長途總費用高於其計劃中指定的費用
總收入	表示公司從該客戶獲得的總收入，計算到上述指定季度末（總費用 - 總退款 + 額外數據費用 + 額外長途費用）
客戶狀態	指示季度末客戶的狀態：已流失、已留下或已加入
客戶流失類別	客戶流失原因的高級類別，在他們離開公司時詢問：態度、競爭對手、不滿意、其他、價格（與流失原因直接相關）
客戶離開原因	客戶離開公司的具體原因，離開公司時被詢問（與流失類別直接相關）
郵遞區號	客戶主要居住地的郵遞區號
人口數	整個郵遞區號地區目前的人口估計

資料分析方法參考說明(部分)

A.探索性分析特徵欄位



B. 移除或修正有問題的特徵

cid	r	f	m	s
				

...

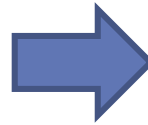
C. 新增分析欄位

cid	r	f	m	s	W	X	Y

...

D. 特徵正規化或相關處理

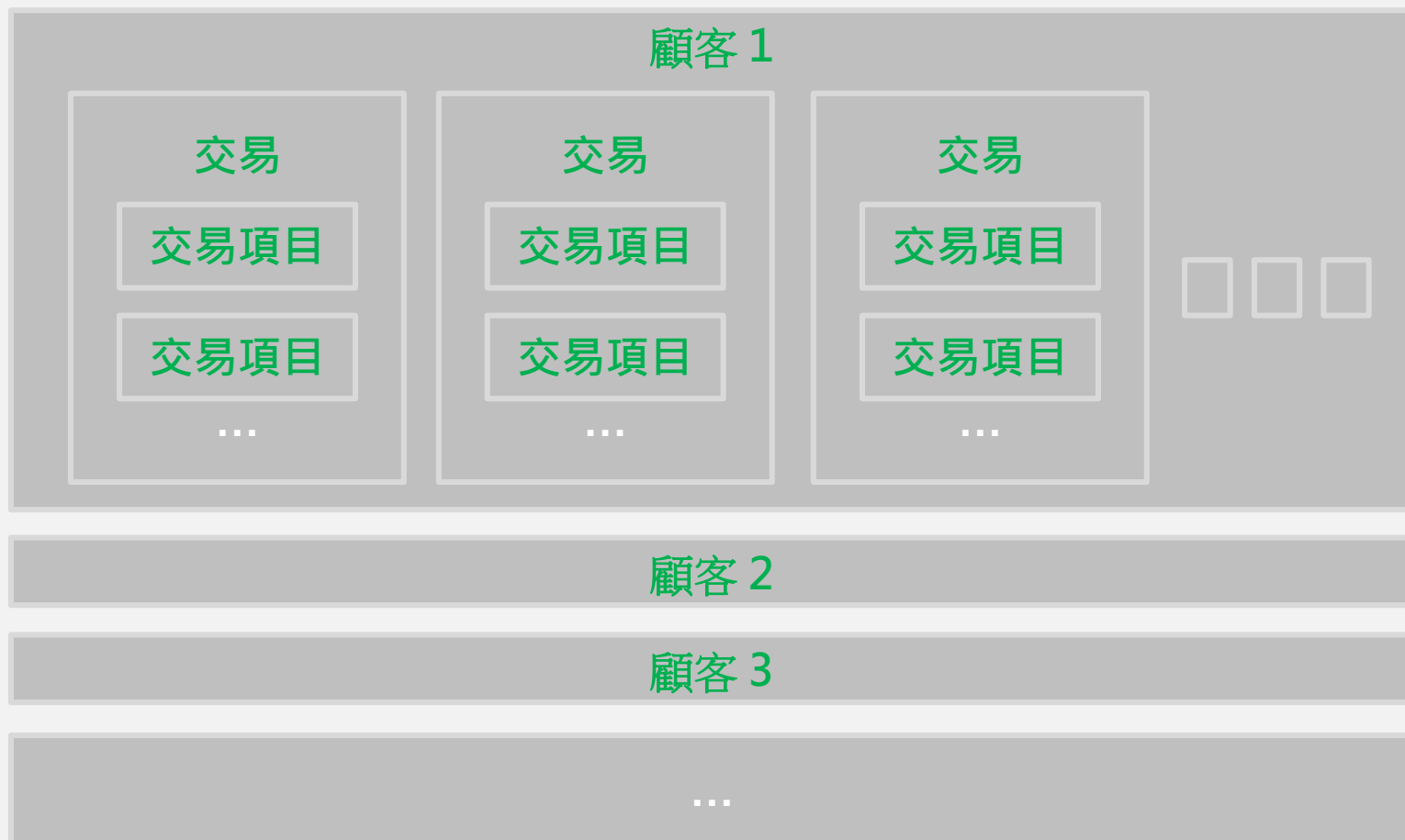
cid	r	f	m	s
		10000		
		50000		
		70200		
		30100		
		10000		



cid	r	f	m	s
		0.0001		
		0.0025		
		0.0030		
		0.0005		
		0.0001		

...

E. 交易資料彙整



F. 資料切割(預測與目標變數)

製作變數 Feature Engineering

1

預測變數
X

目標變數
Y

[illegible]

分割資料

Data Splitting

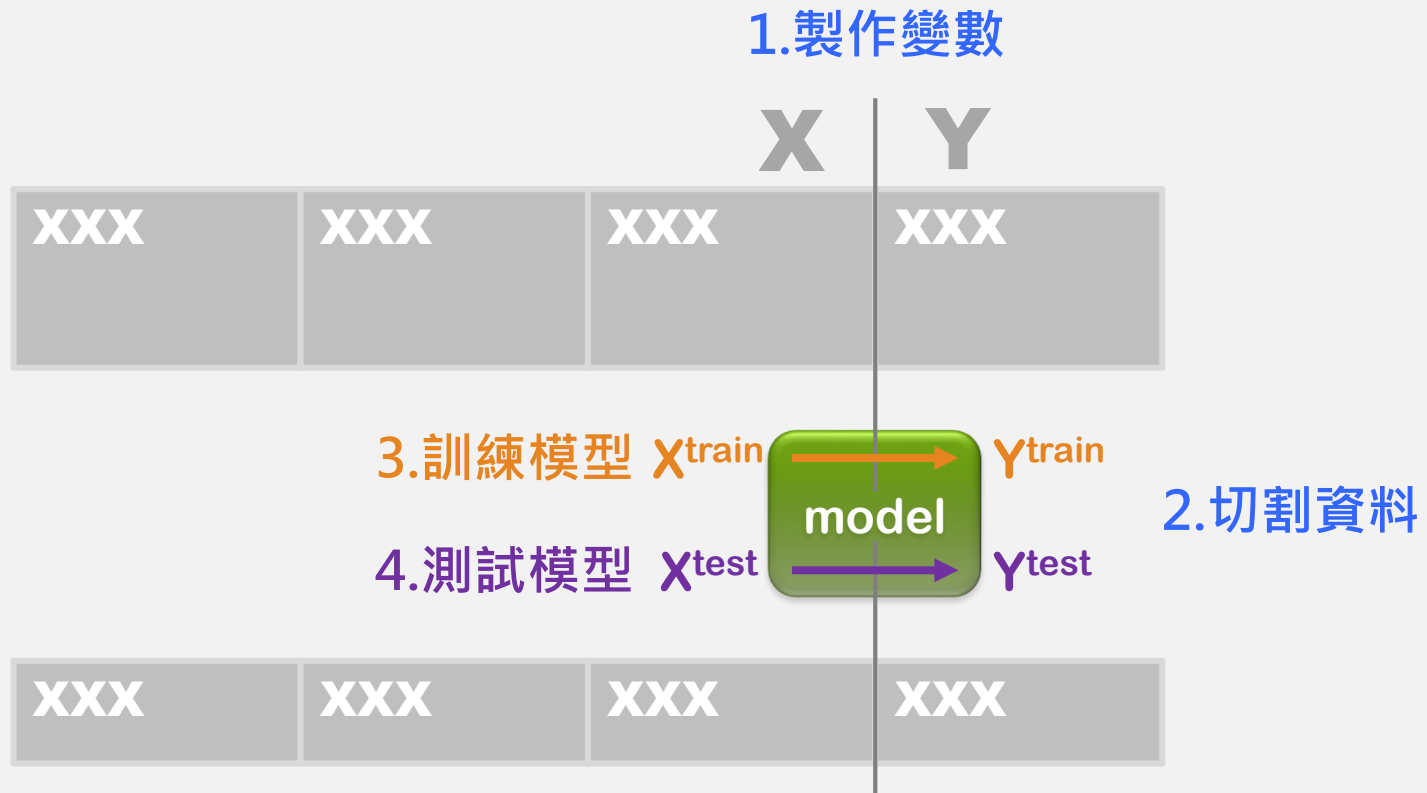
2

訓練資料 TR

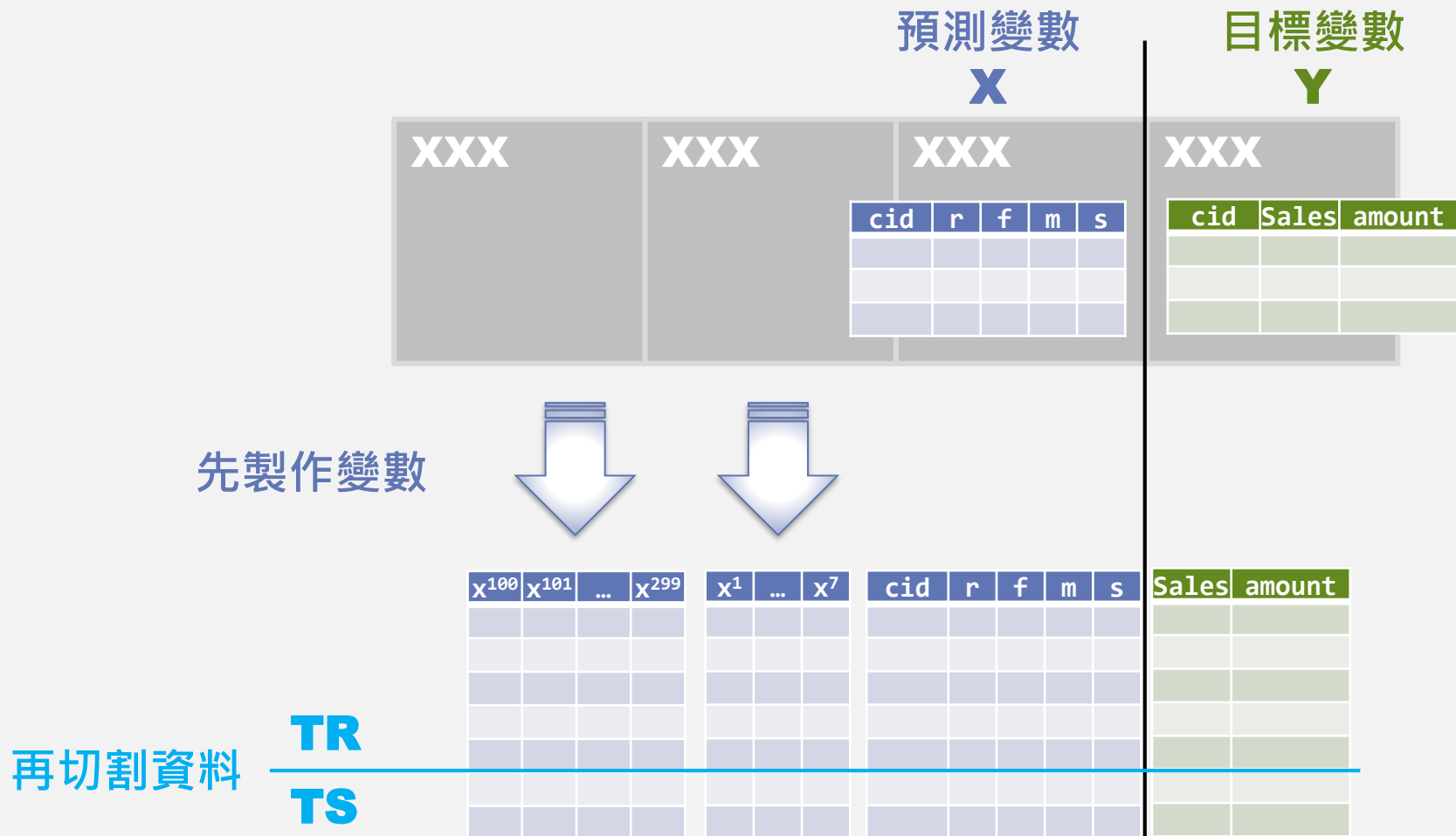
測試資料 TS

[illegible]

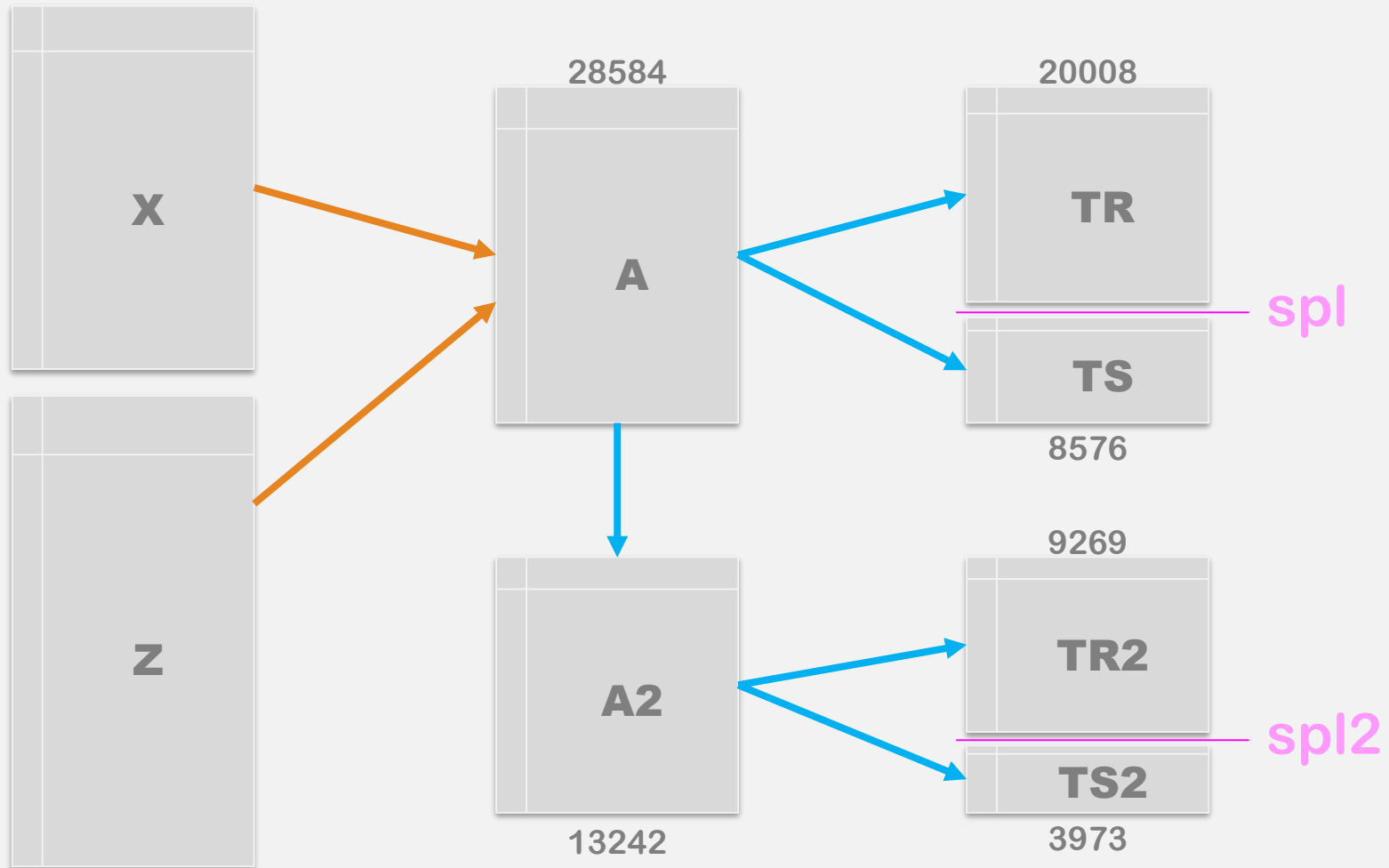
G. 建立模型



H. 製作變數、改進模型



Feature Engineering & Data Splitting



I. 進行預測及評估



The End