



113年國道智慧交通管理創意競賽



國道1號里程10-99K車禍回堵里程預測

我的獎金~~15萬!



目錄

01	背景介紹	1
02	文獻回顧	2
03	研究方法	3
	• 資料來源	• 模型開發
	• 特徵擷取	• 實際應用模型流程
	• 變數篩選	• 研究結果討論
04	實務應用與可行性	36

背景介紹



WHY 事故回堵里程？

- 現今1968、資訊可變標誌、即時路況平台多提供即時資訊

缺乏未來短期路況變化預測

使得回堵更加嚴重導致政府人員調度困難！

WHY 10-99K ?

- 根據政府公布國道易雍塞路段數據

1. 經過許多人口密集區，如：台北內湖、新北林口、桃園南崁、新竹工業園區
2. 連結許多易雍塞路段與重要匝道

文獻回顧



相關文獻回顧

A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data

作者: Amirhossein Abdi, Seyedehsan Seyedabrihami, and Steve O'Hern

目標: 該道路發生事件後，需多久時間回到原車流水準

方法: 二階段建模

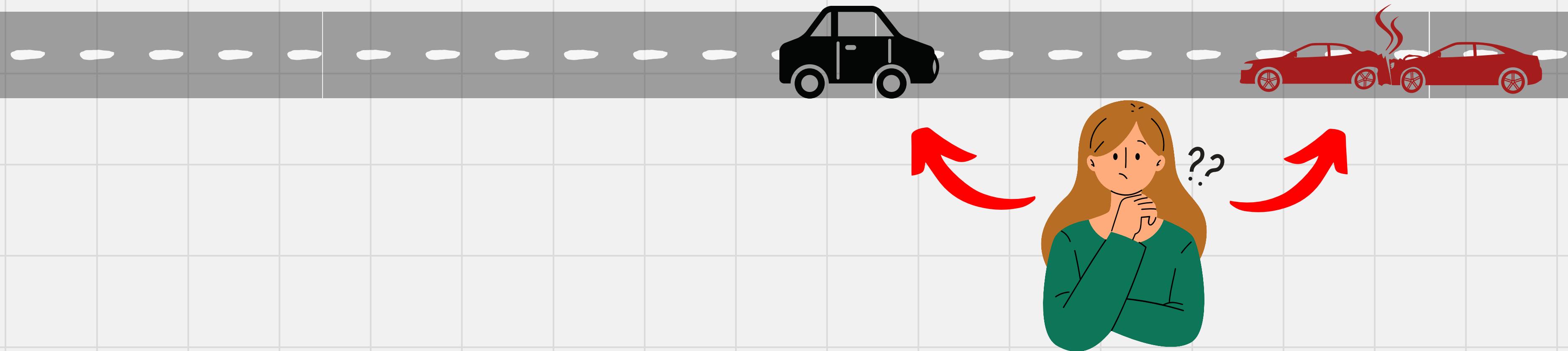
第一階段(警察抵達前)

- 類別型預測，預測事故後車流等級變化
- XGBoost 準確率約 0.73-0.83
- 特徵以交通即時資訊較重要

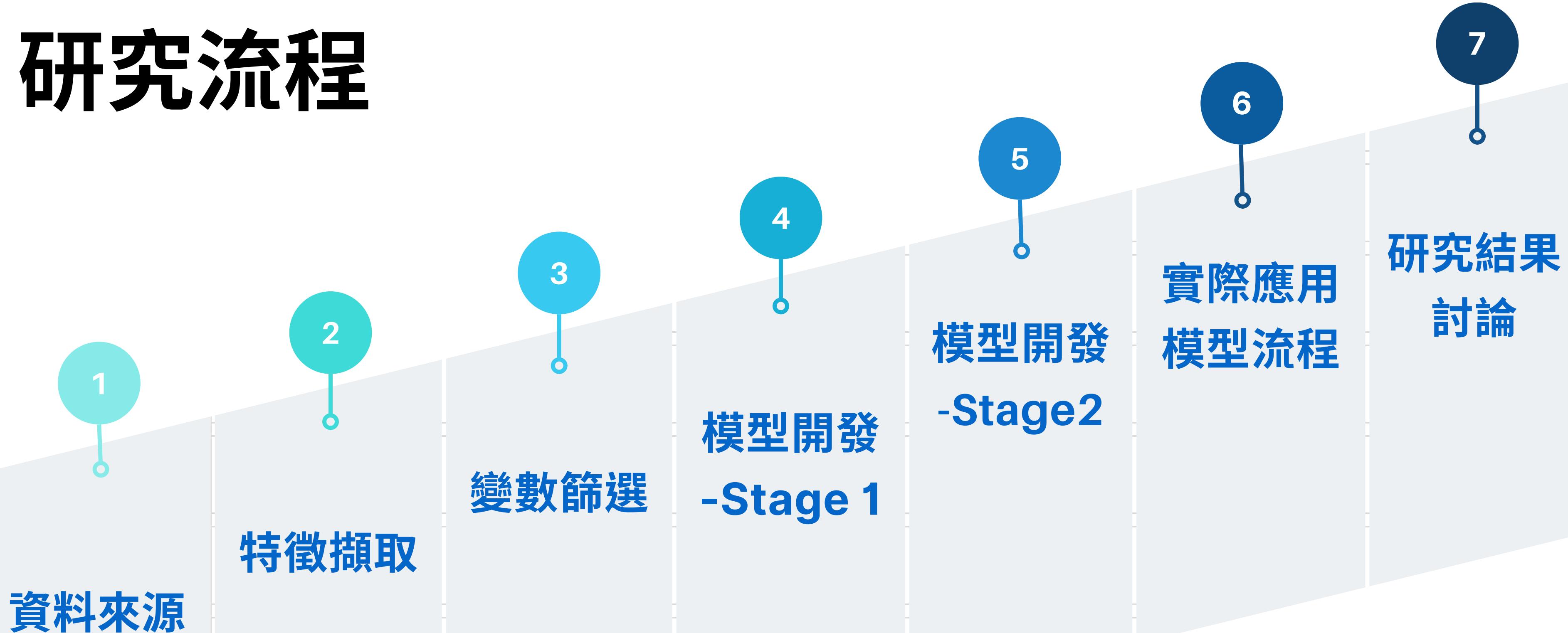
第二階段(警察抵達並處理後)

- XGBoost 連續型預測
- 警察抵達後若由第一階段模型可能等級更糟，則預測仍需多少時間回到原等級
- RMSE: 8.86分鐘
- MAPE: 10.44%
- 特徵以事故因子資訊較重要

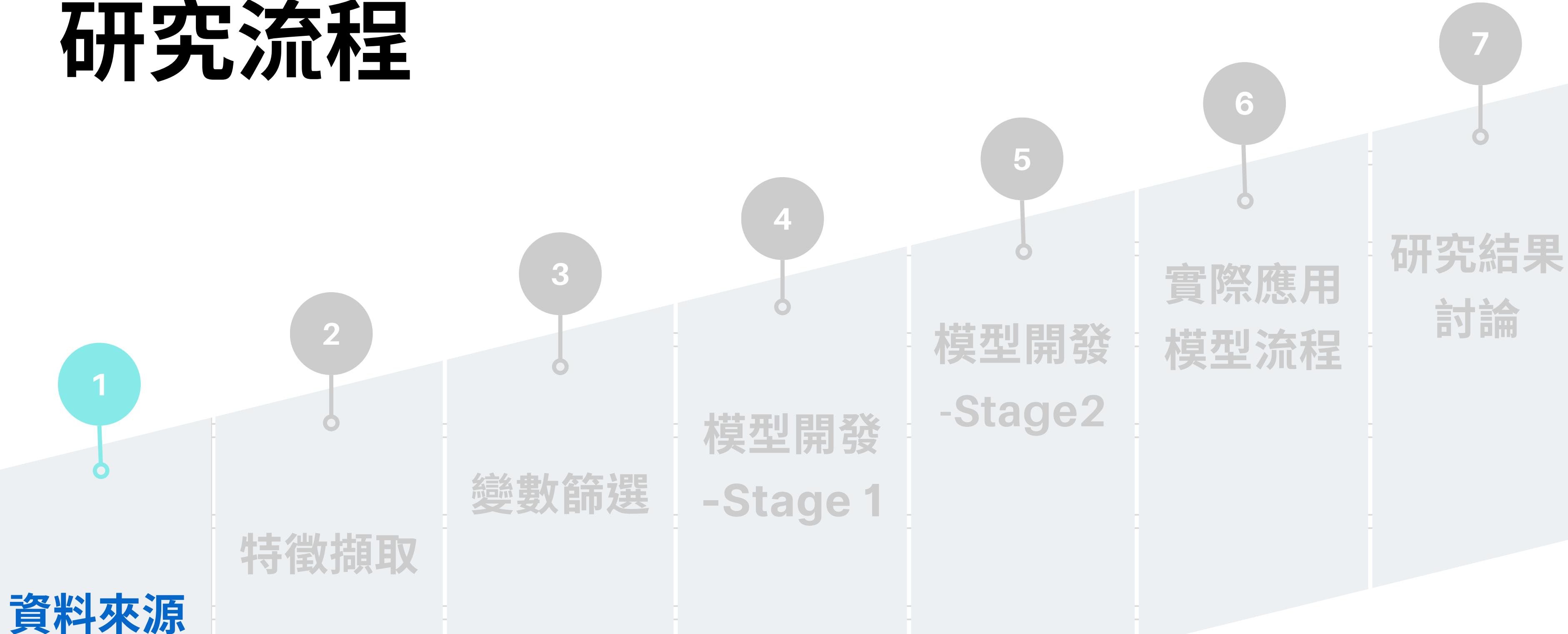
研究方法



研究流程



研究流程



資料來源

- 天氣資訊：timeanddate 網站



<https://www.timeanddate.com/>

- 交通即時資訊：高工局交通資料庫



交通部高速公路局
FREEWAY BUREAU, MOTC

交通資料庫

- Etag靜態資訊(v2.0)
- 各類車種通行量統計各類車種(M03A)
- 各旅次路徑原始資料(M06A)

<https://tisvcloud.freeway.gov.tw/>

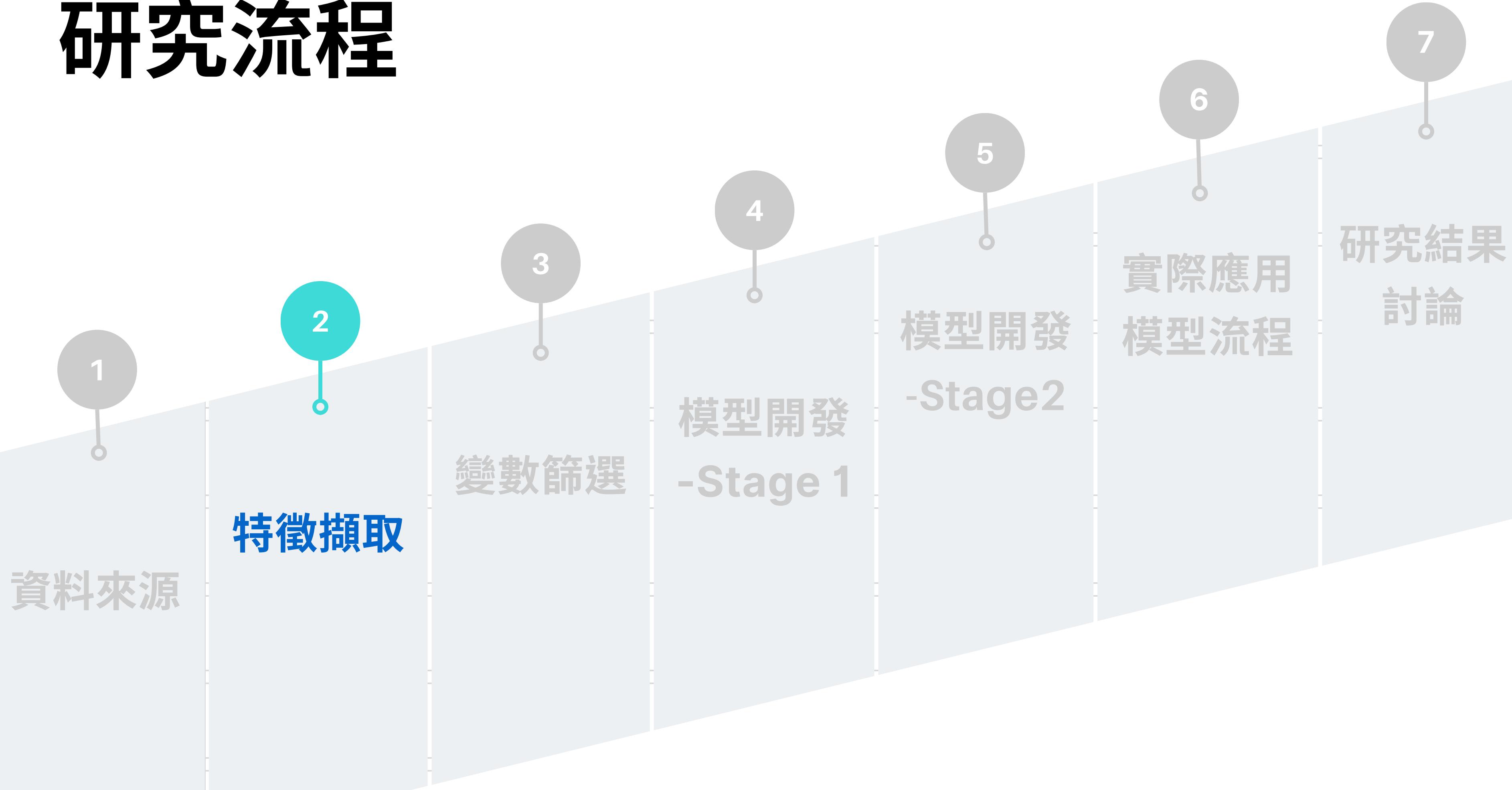
- 事故資訊：競賽官網提供2023年1月至10月國道事故簡訊



- 施工資訊：競賽官網提供2023年1月至10月施工路段資料

<https://freeway2024.tw/links#links> 3

研究流程



特徵擷取-天氣資訊



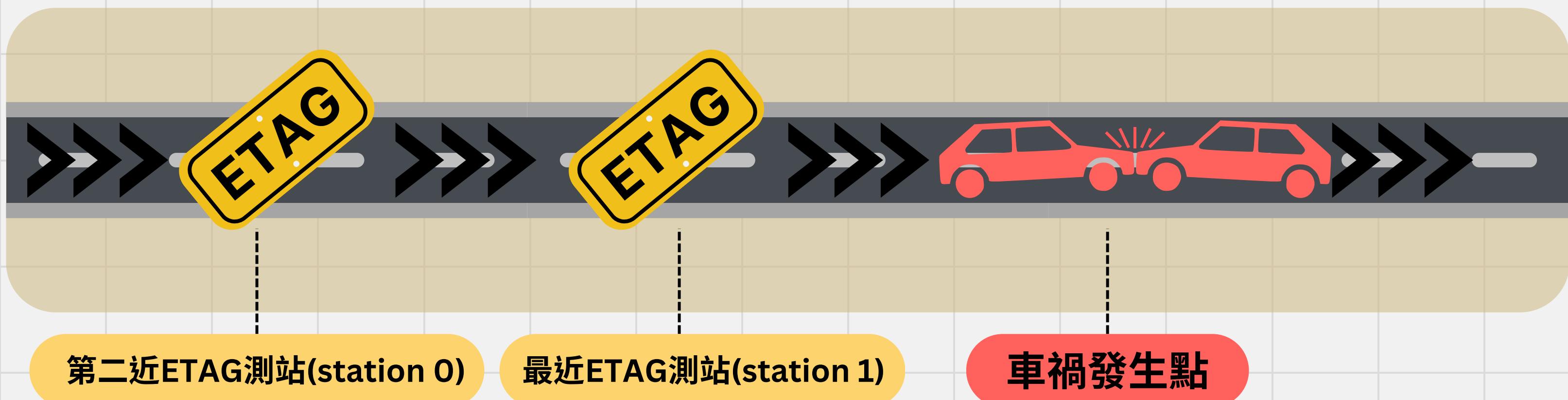
- timeanddate網站有對台灣各地區、每小時的詳細氣象資訊
- 提供該地區**溫度**、**氣壓**、**風速**、**濕度**等資訊

Conditions			Comfort			
Time	Temp	Weather	Wind	Humidity	Barometer	Visibility
00:00 Sun, 1 Jan	17 °C	Light rain. Passing clouds.	20 km/h	— 94%	1024 mbar	9 km
01:00	17 °C	Light rain. Passing clouds.	24 km/h	— 94%	1024 mbar	6 km
02:00	17 °C	Light rain. Passing clouds.	26 km/h	— 94%	1023 mbar	7 km
03:00	18 °C	Light rain. Passing clouds.	19 km/h	— 94%	1023 mbar	8 km
04:00	18 °C	Light rain. Passing clouds.	22 km/h	— 94%	1023 mbar	N/A
05:00	18 °C	Light rain. Passing clouds.	22 km/h	— 88%	1023 mbar	9 km
06:00	18 °C	Light rain. Passing clouds.	26 km/h	— 94%	1023 mbar	N/A
06:30	18 °C	Passing clouds.	22 km/h	— 94%	1024 mbar	N/A

- 利用**網路爬蟲**蒐集各事件最近時間的天氣資訊
- 缺點是無降雨量，且沒測站的詳細資訊，只能大致由里程分類至可查詢地區

特徵擷取-交通即時資訊

- 此資訊目的是協助評估事故發生前的道路狀況。如:車流、車速
- 測站資料及定義:
 - 測站資料從高工局資料庫靜態資訊(v2.0)提取
 - 定義station0及station1



特徵擷取-交通即時資訊

事故前十分鐘的車流量及大型車比例

- 此特徵主要從高公局M03A提取，資料是以每五分鐘為時間區段
如：檔案2023-08-17 06:00代表 2023-08-17 06:00 – 06:05 的總車流

1. 尋找該事件的 Station 1 (若 station 1 無資料則選 Station 0)
2. 提取該事件前十分鐘**總車流量**並計算**大型車比例**

總車流量 = 小客車量 + 小貨車量 + 大客車量 + 大貨車量 + 聯結車量

大型車比例 = (大客車量 + 大貨車量 + 聯結車量) / 總車流量

特徵擷取-交通即時資訊

事故後兩分鐘的車流量

- 此特徵主要從高公局M03A提取，資料是以每五分鐘為時間區段，以**權重平均**擷取。
- 事故在03:32發生，提取03:30 – 03:35車流量資料並取其中2/5
- 事故在03:29發生，提取03:25 – 03:30及03:30 – 03:35車流量資料並各取其中1/5
- 事故在03:35發生，提取03:35 – 03:40車流量資料並取其中2/5

特徵擷取-交通即時資訊

事故前十分鐘平均車速

- 此特徵主要從高公局M06A提取，此資料追蹤車輛經過各門道的時間點
- 針對單一事件篩出其前十分鐘經過Station 1 的車輛，假設共n輛。

單一車輛車速計算：

speed : 該車當時車速

$$speed_j = \frac{\text{測站里程差(km)}}{\text{時間差(s)}}, j = 1 \dots n$$



單一事件事故前十分鐘平均車速計算：

Pre_Average_CarSpeed : 該事件前十分鐘路況車速，共5702條事件

$$\text{Pre_AverageCarSpeed}_i = \frac{\sum_{j=1}^n speed_j}{n}, i = 1 \dots 5702$$

特徵擷取-簡訊

- 此類特徵由競賽網站提供的112年1-10月與113年1-2月交通事故簡訊通報資料擷取，以資料表中的"簡訊內容"拆分出所需特徵。

直接擷取：

時間類別、方向、里程、回堵里程、死亡、受傷、事故波及車道，其中事故波及車道並非來自簡訊內容，而是使用原始表格內的欄位，包含五個車道、內外路肩以及匝道。

間接擷取：

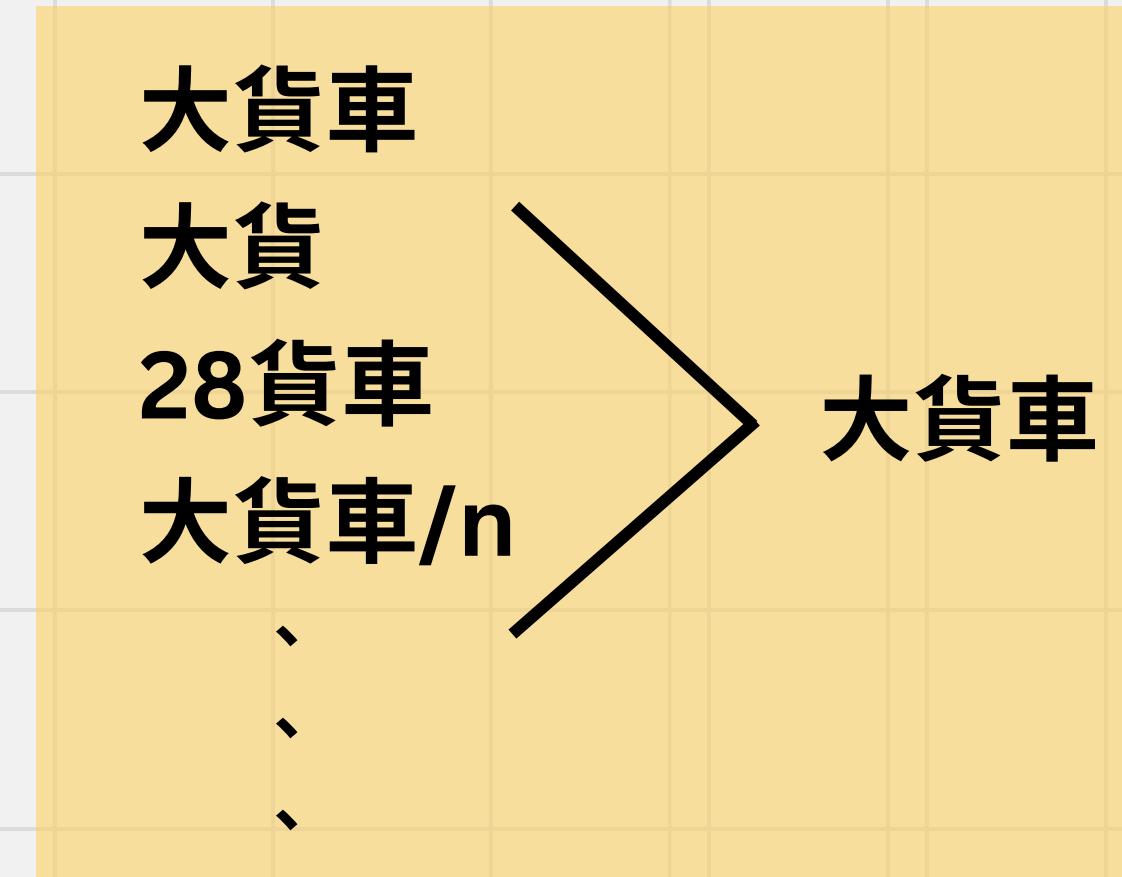
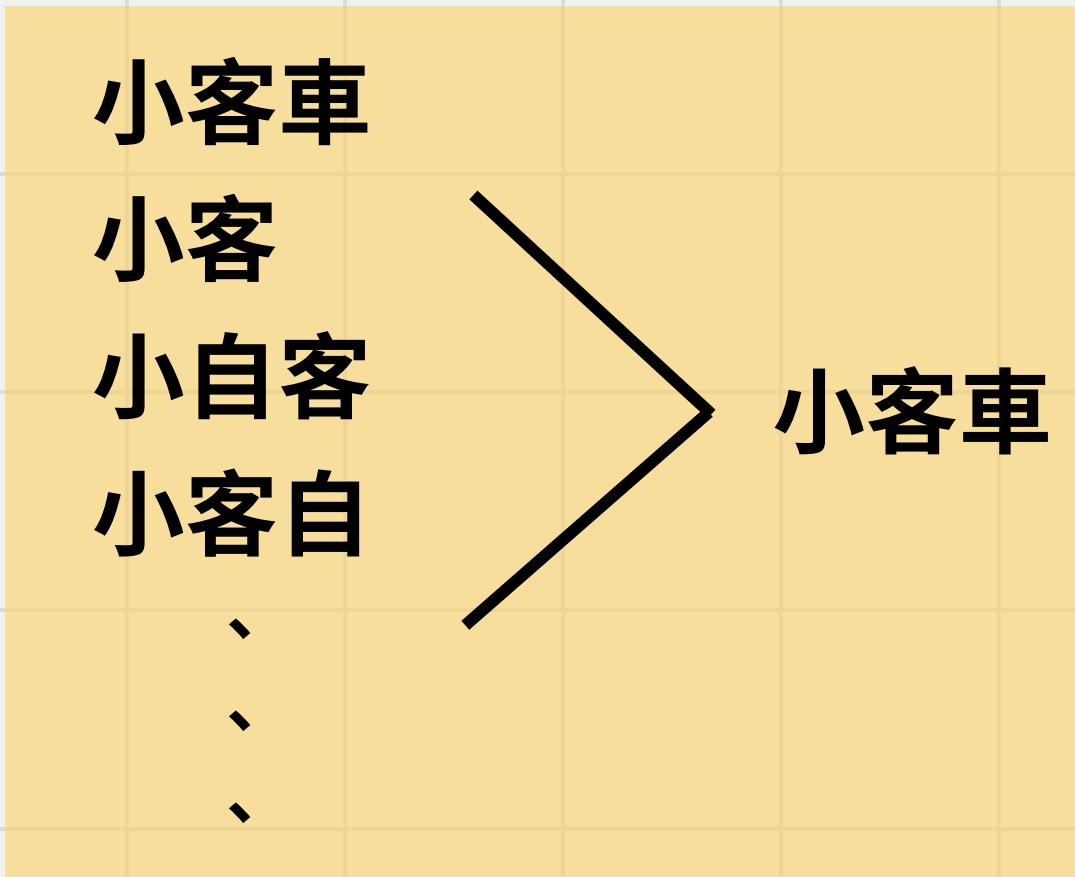
日期類別 - 依擷取日期分成平日、例假日與國定假日。

處理分鐘 - 用事故排解完成時間與事故發生時間相減而得。

特徵擷取-簡訊

車輛相關欄位

- 此特徵來自簡訊擷取之受波及車輛，獨立紀錄後再進行分類。



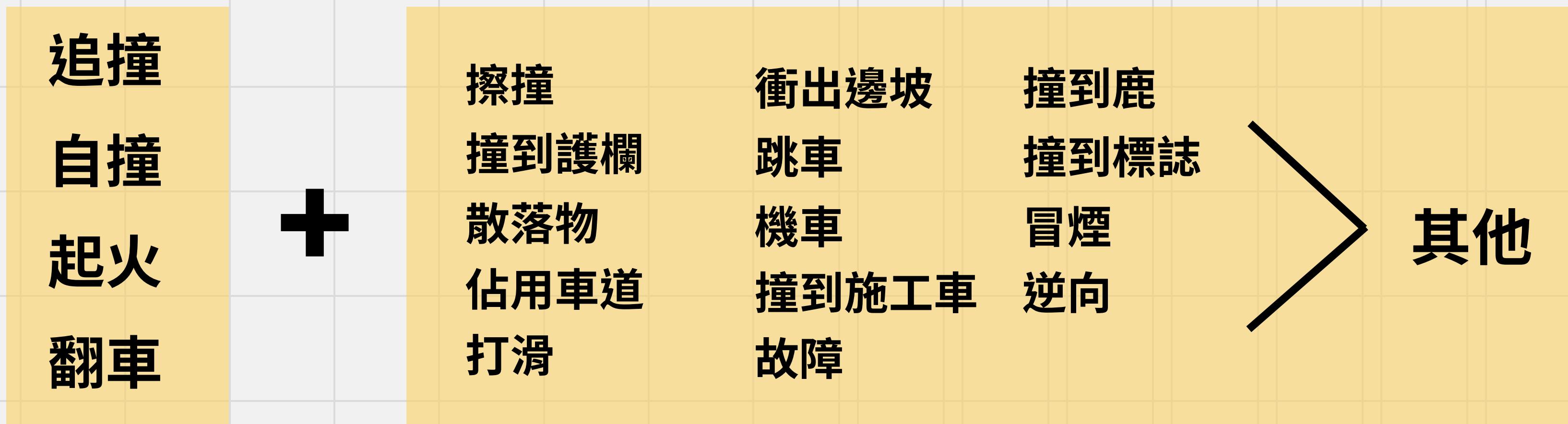
以此類推

歸類出工程用車輛、小客車、中大型客車、中小型貨車、大貨車及聯結車、其他車輛等六種，再按每筆事故計算參與的車類數量加總紀錄，最後再全部加總計算出肇事車輛總數。

特徵擷取-簡訊

事故原因相關欄位

- 此特徵來自簡訊擷取之肇事原因，大致分為五類。



因為同一事故簡訊可能同時包含多個原因，如散落物和翻車，因此當同一事故的原因中包含追撞、自撞、起火、翻車等四個原因時會優先記錄，剩下皆為其他原因。

特徵擷取-施工

- 此特徵由競賽網站提供的112年1-10月與113年1-2月道路施工路段資料擷取，先篩選出此次實驗路段的施工資料，並記錄各施工佔用的車道總數。

事故與施工案件配對原則



: 事故位置前後300公尺

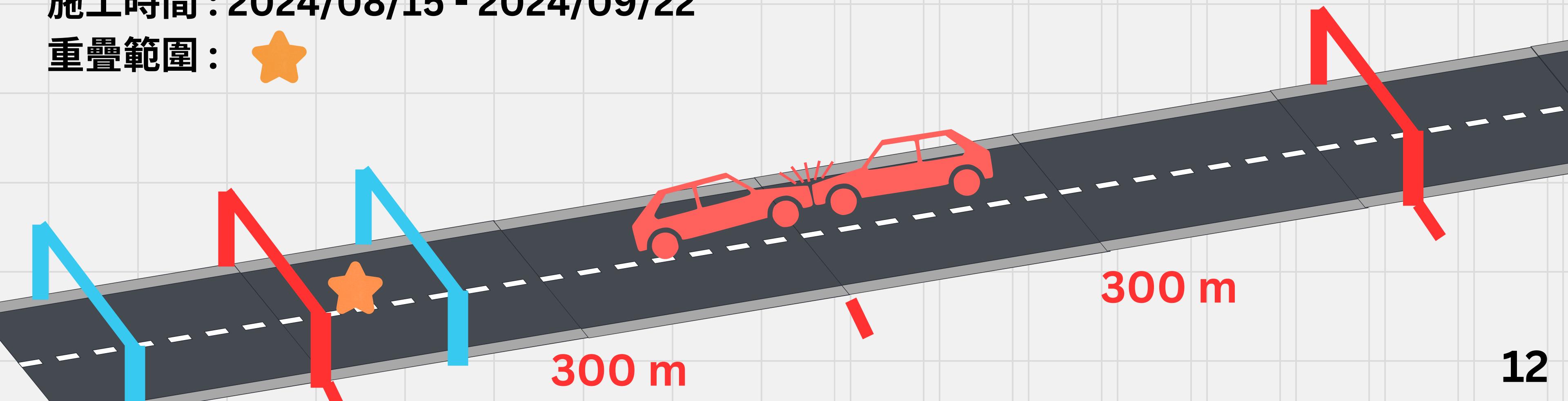


: 施工路段

事故時間 : 2024/08/17 08:55

施工時間 : 2024/08/15 - 2024/09/22

重疊範圍 :



300 m

300 m

特徵擷取-所有變數

施工因子

- 施工占用車道

氣候因子

- 溫度、風速、濕度、氣壓

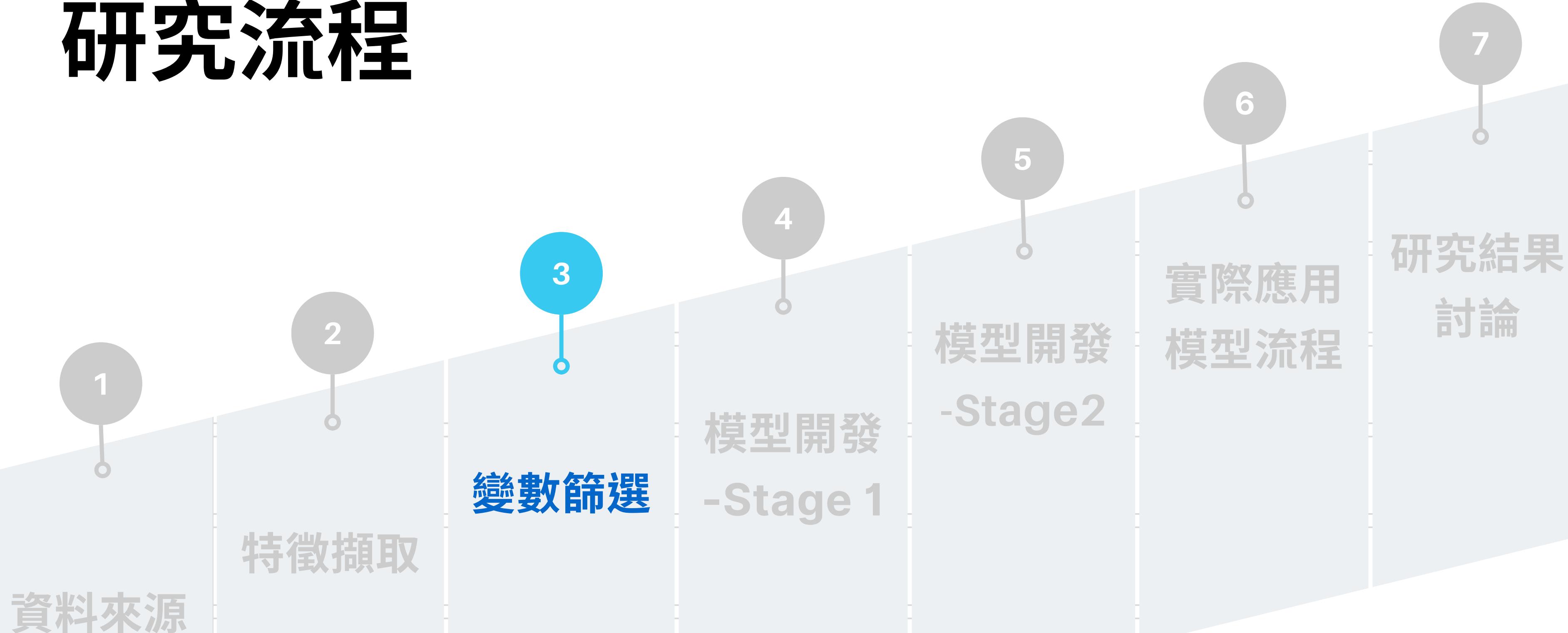
事故因子

- 事故前大型車比例
- 事故前平均車速
- 事故前車流量
- 事故後車流量

交通即時因子

- 事故概述（時間、傷亡...）
- 事故波及車道狀況
- 事故波及車輛種類
- 事故原因

研究流程



變數篩選

所有變數

Lasso算法

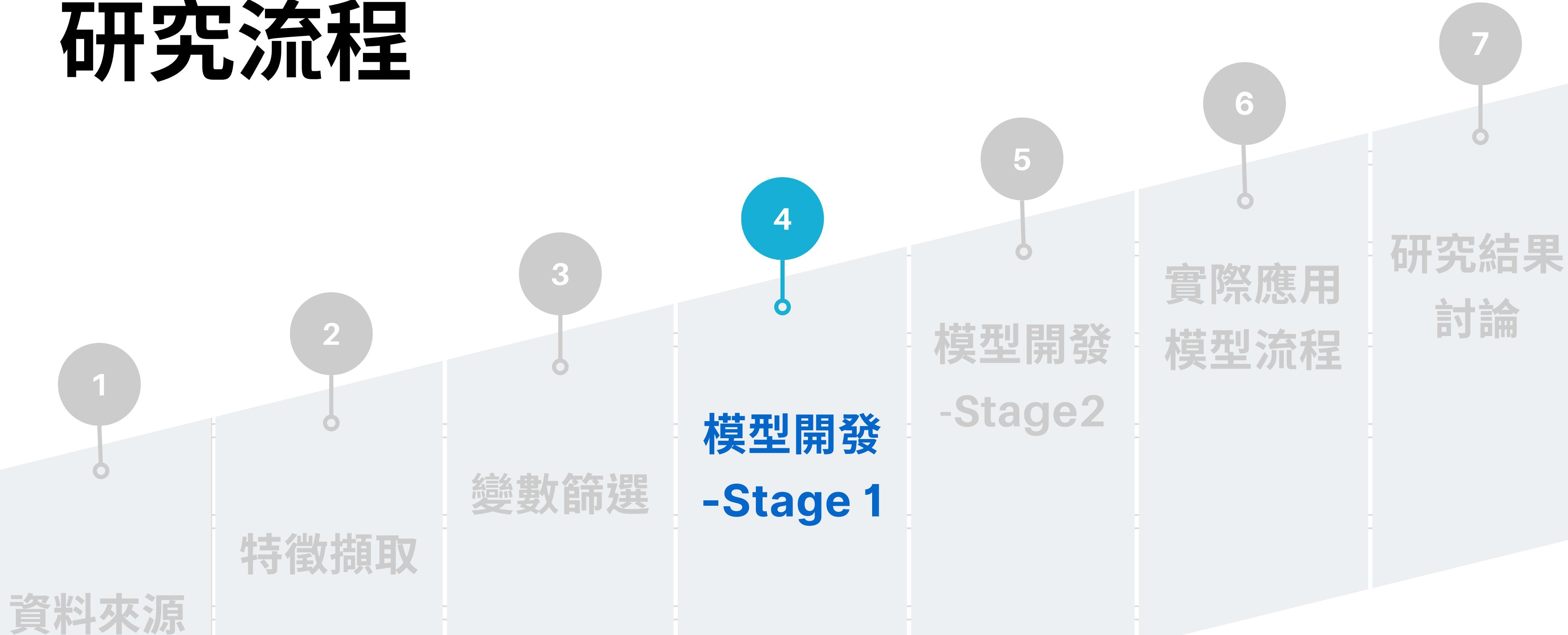
Backward
Selection

卡方檢定
&相關係數

- 一般欄位將受傷與死亡合併為傷亡變數。
- 車輛欄位刪除小客車，保留事故車輛數，並將其他種類車輛合併為大型車數量。
- 事故原因欄位保留翻覆，並將除翻覆和追撞外的原因合併進追撞中。
- 天氣欄位中刪除溫度、濕度、氣壓計三個變數。



研究流程



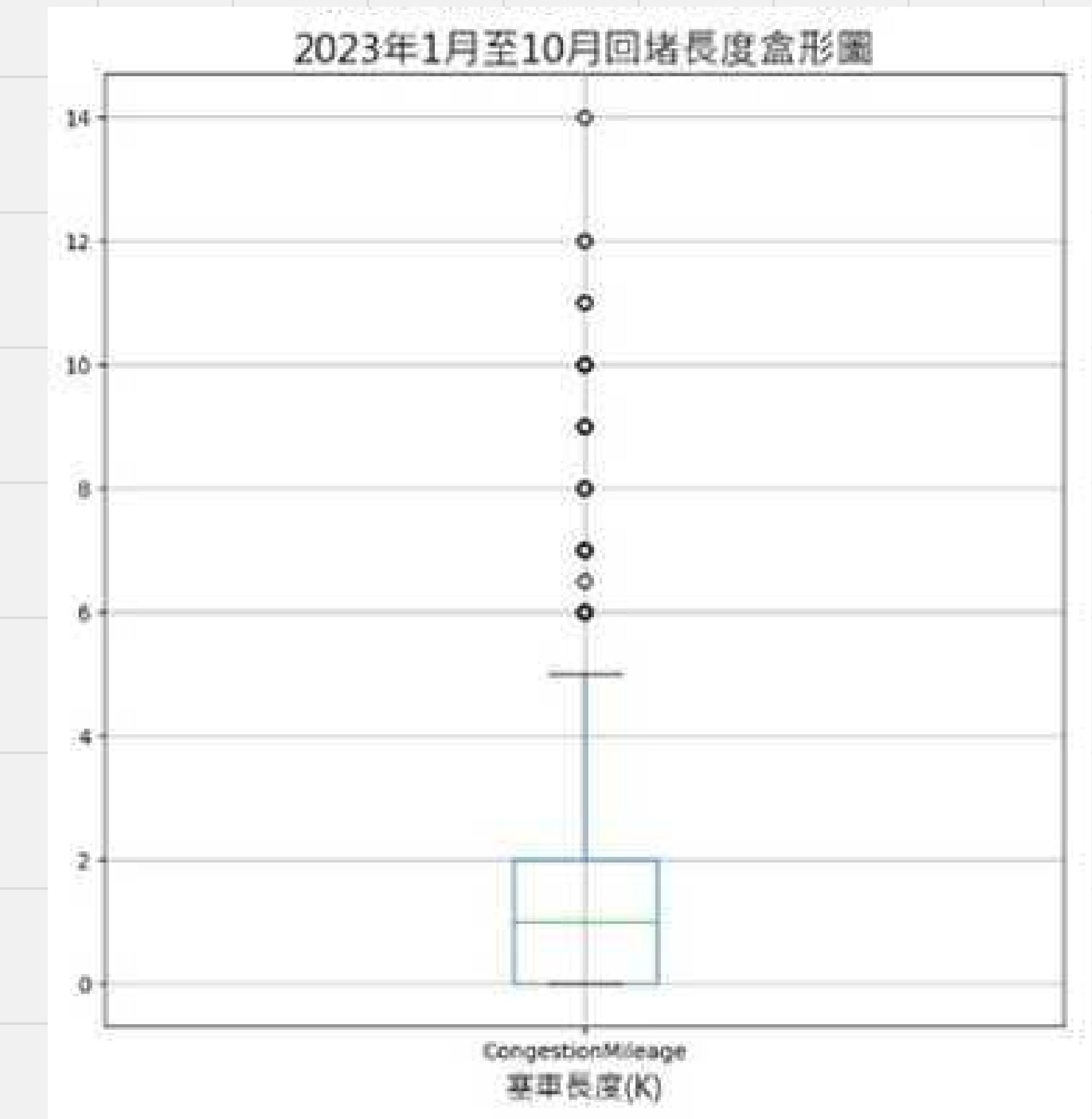


模型開發-初步構想

- 回堵公里平均約 1.13K
- 全距到14K

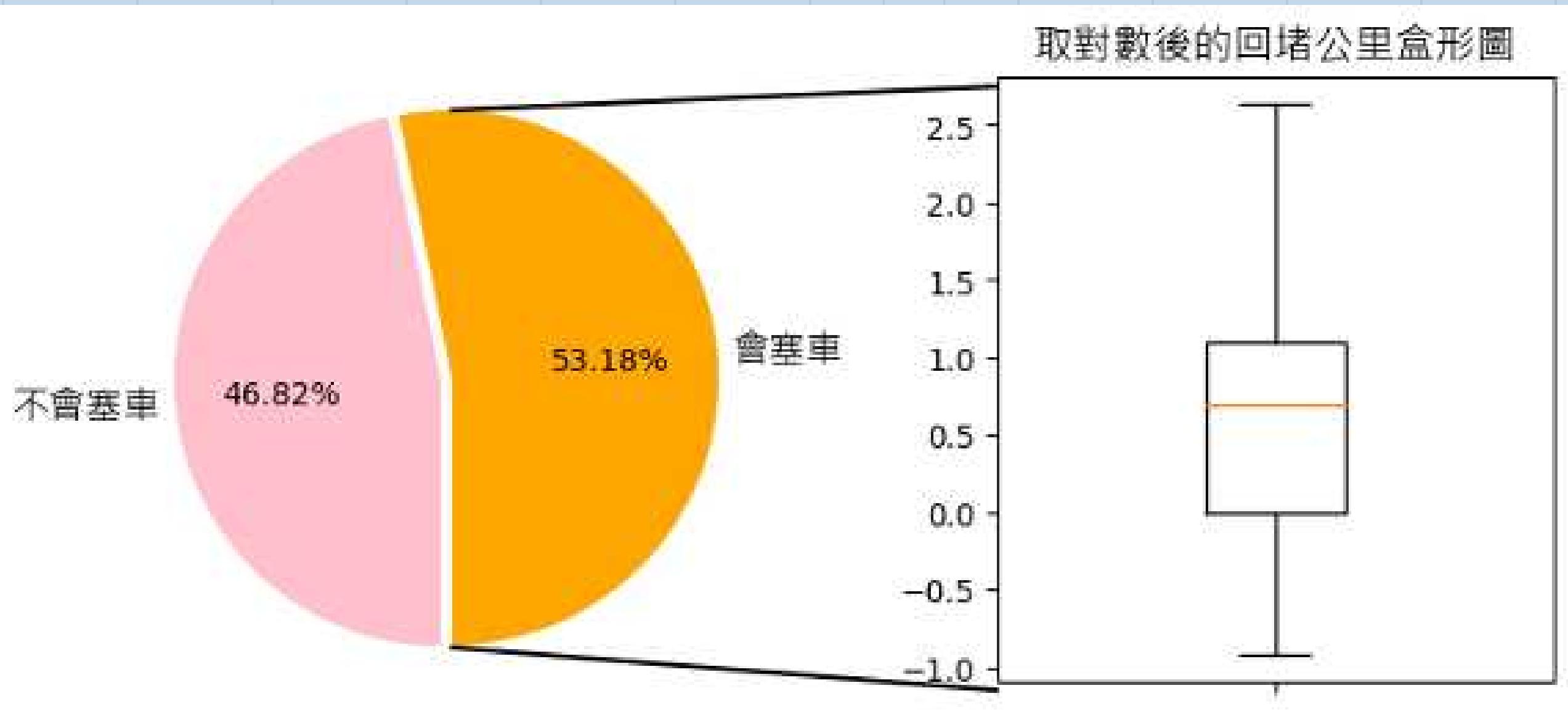


反應變數變異程度偏大，
增加預測難度





模型開發-初步構想



- 是否塞車的資料平衡
- 對有回堵事件的里程取對數以縮小全距、平滑數據



模型開發-初步構想



參考Amirhossein Abdi等人之文獻

Two sequential framework 建模結構

First stage

- 針對事件是否回堵進行預測

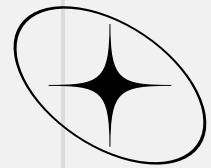
是

否

Second stage

- 針對 $\log(\text{回堵長度})$ 建模
- 預測其回堵長度

該事件回堵長度為0



第一階段模型

目標

預測「Congestion」(是否回堵)

本階段使用之訓練集

112 年 1~10 月所有車禍的數據

本階段使用之測試集

113 年 1~2 月所有車禍之數據



第一階段模型選擇指標

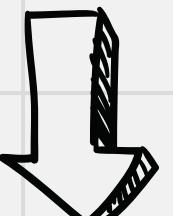
		預測	
		是	否
真實	是	TP	FN
	否	FP	TN

混淆矩陣

第一階段模型評估根據混淆矩陣參考以下指標：

準確率(Accuracy)

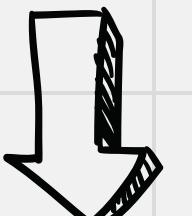
正確識別是否回堵的能力



$$\frac{TP+TN}{TP+TN+FP+FN}$$

召回值(Recall)

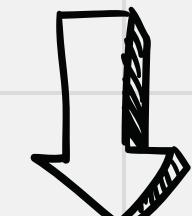
識別實際上有堵塞狀況的能力



$$\frac{TP}{TP+FN}$$

精確度(Precision)

正確識別會回堵的能力



$$\frac{TP}{TP+FP}$$

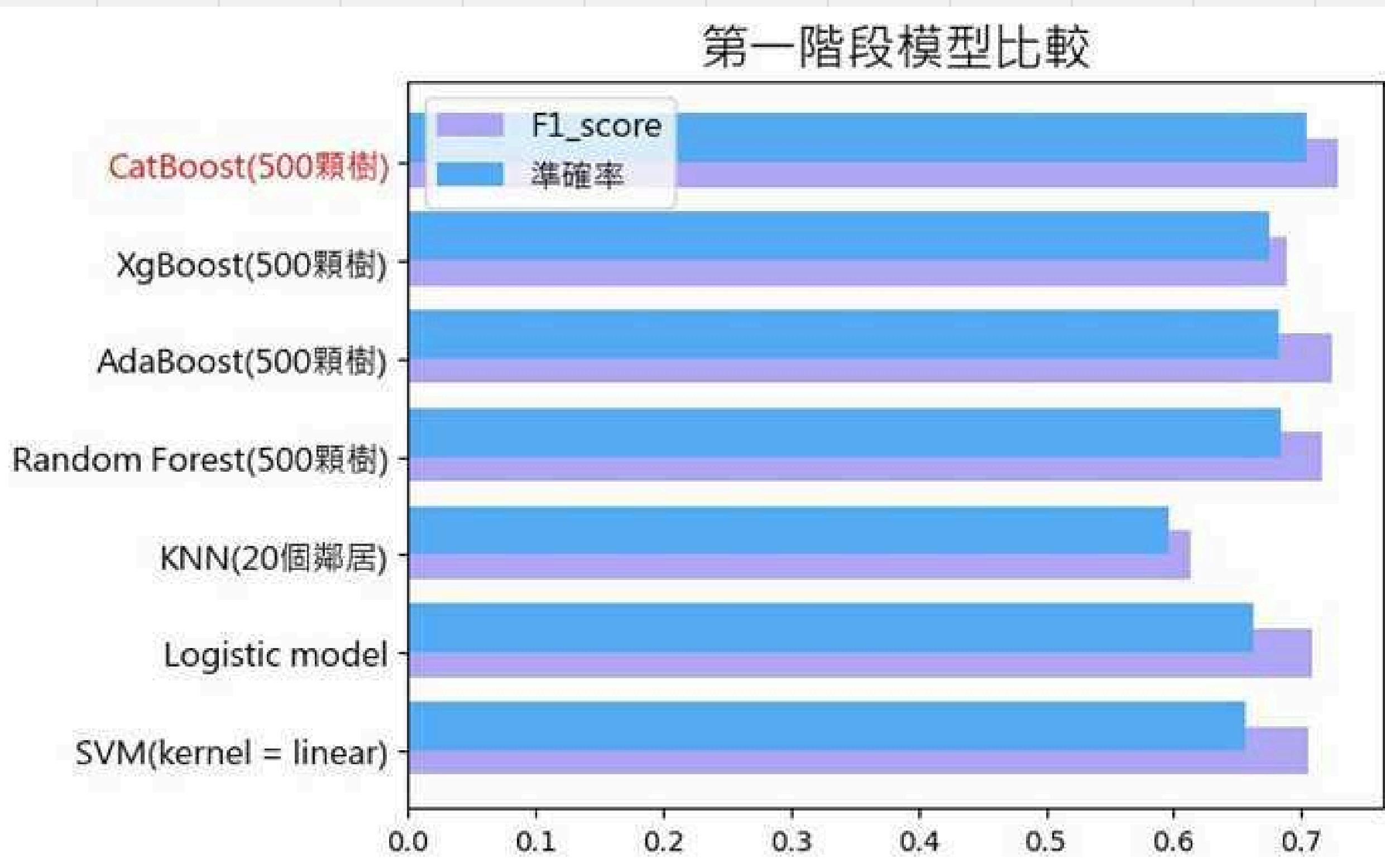
F-Score

精確度與召回值的調和平均數指標

- F1-Score: 同樣注重精確度及召回值
- F2-Score: 較重視召回值



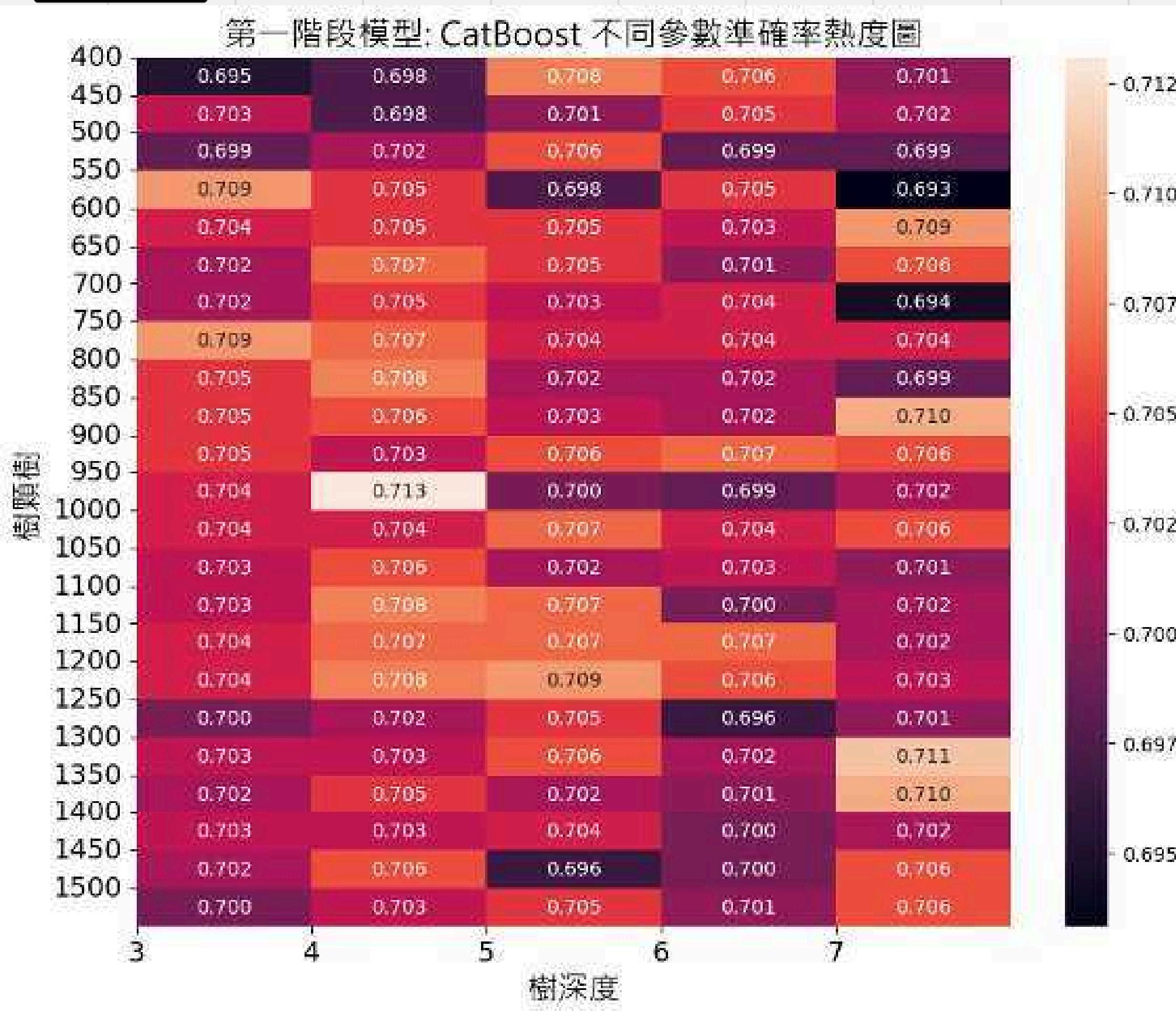
第一階段模型選擇



- 樹狀模型較佳，其中以 **CatBoost** 表現較好，故第一階段選用 CatBoost 模型。
- **WHY CatBoost ?**
 1. 彰顯類別變數內部的順序性及對目標變數的相關性。
 2. 減少類別資料維度
 3. 樹狀結構在捕捉交互作用中較靈活。



第一階段模型參數選擇

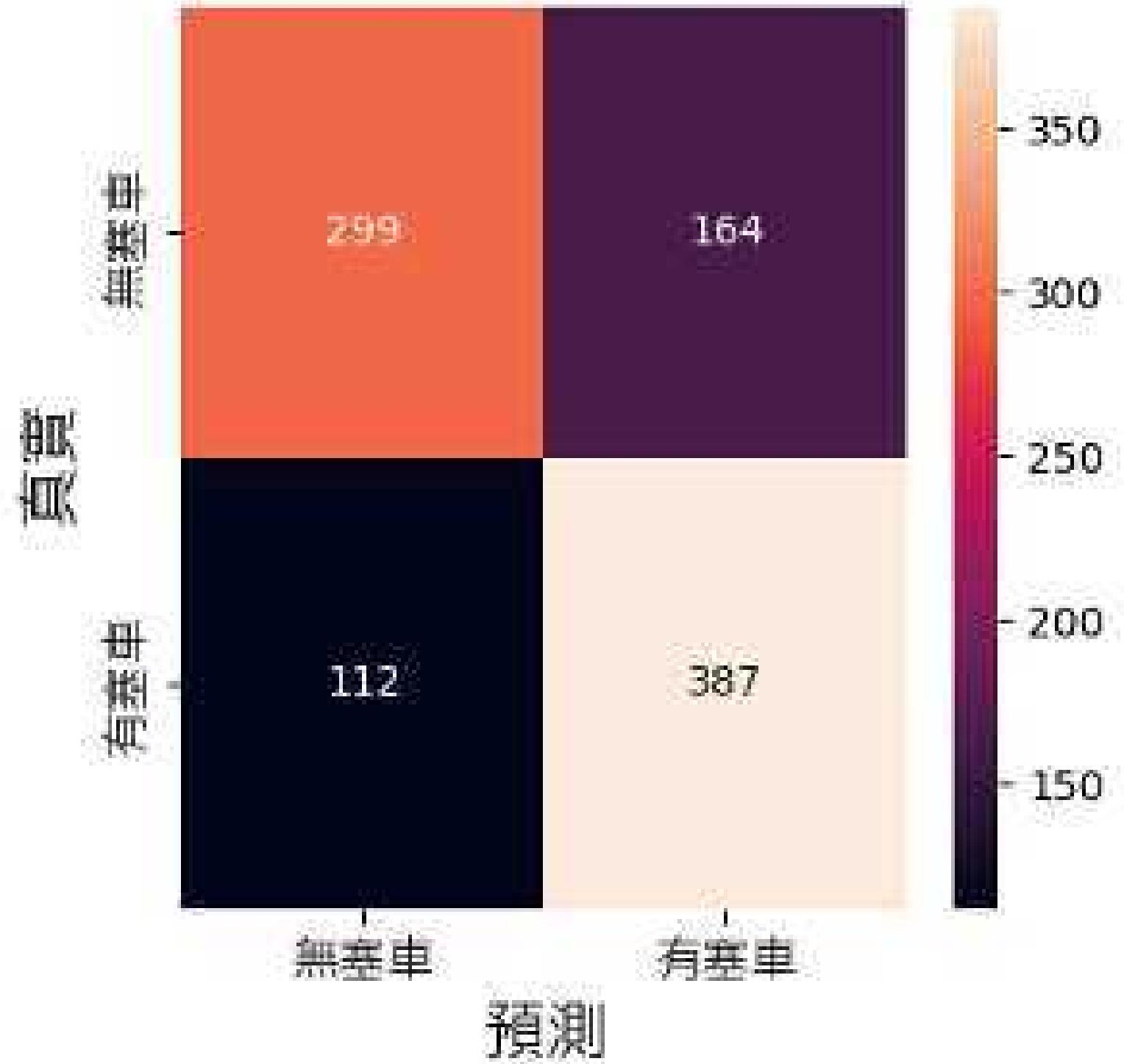


- 對CatBoost做不同參數組合嘗試，其中以950顆樹且深度為4表現最好，即熱度圖中白底處，準確率約0.713

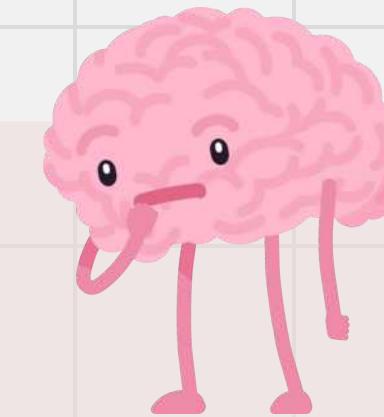


第一階段模型調整

CatBoost(950顆樹、深度為4)之混淆矩陣

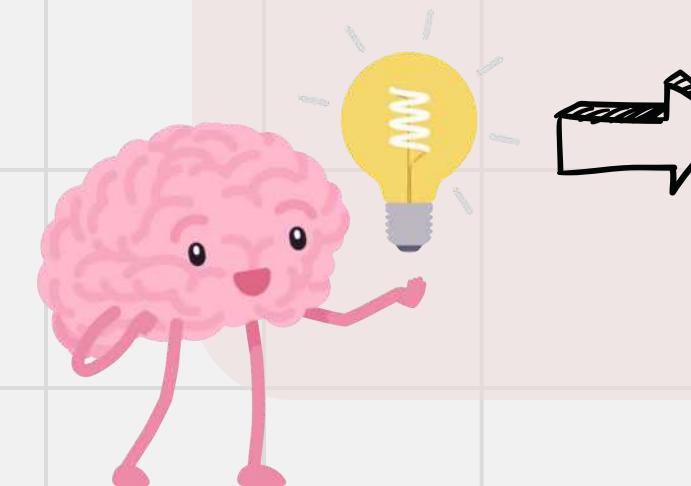


模型結果反思及調整

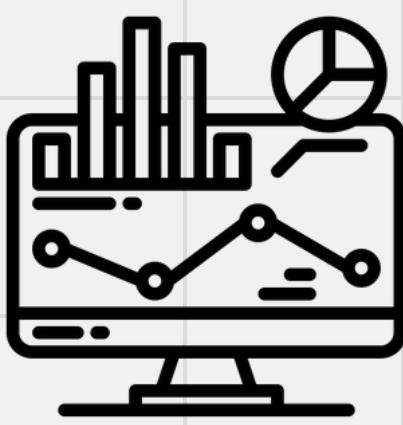


以本次研究目標的FN及FP而言，我們認為FN所付出的成本會較大，即此事件會造成回堵但卻預測不會回堵，容易引起民怨及人員調度混亂。

- 調整模型，使其盡可能網羅會塞車的事件



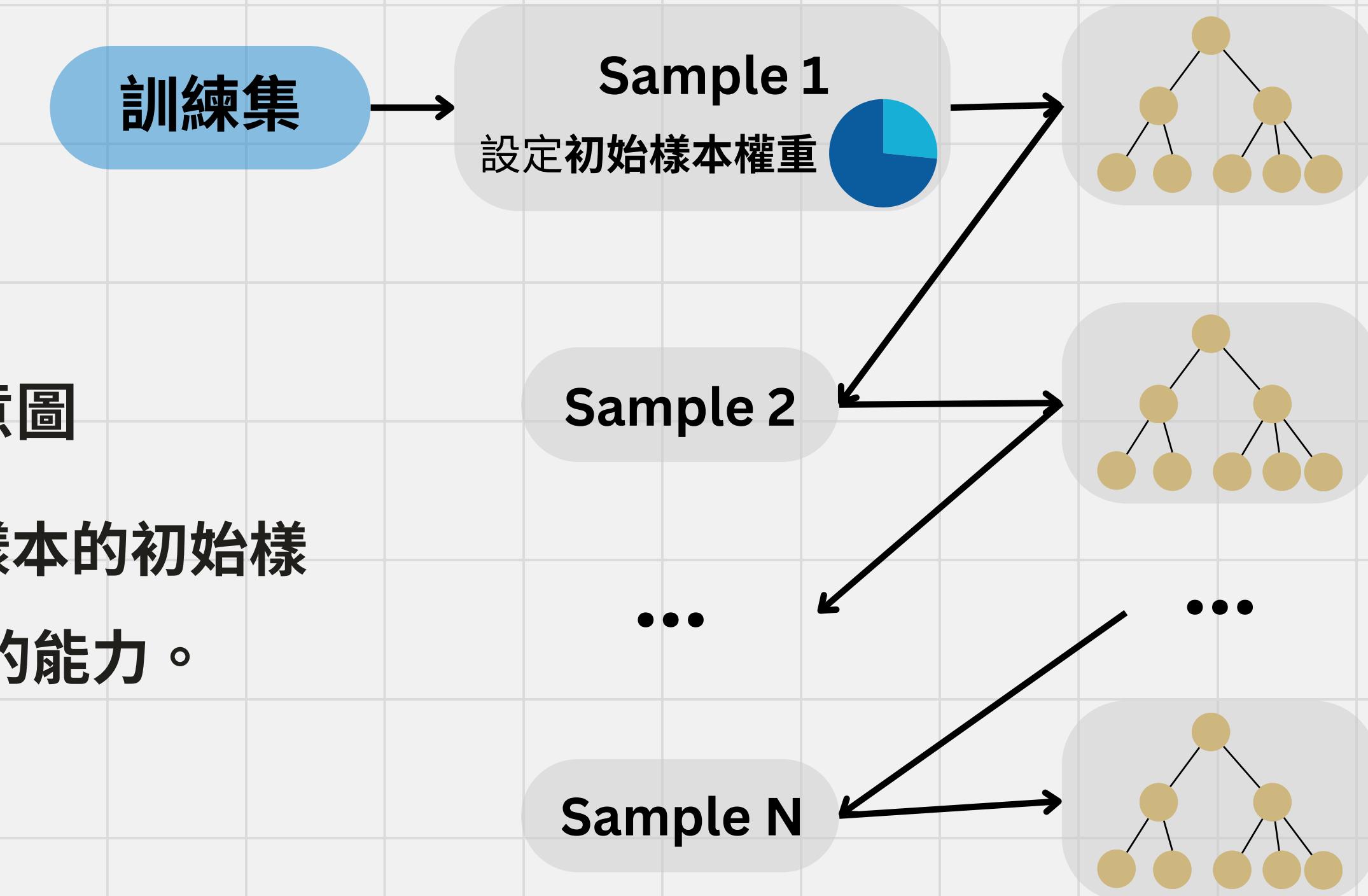
提高召回率(Recall)



CatBoost模型架構及樣本比例

- 右圖為CatBoostClassifier的示意圖

利用其中`class_weights`方法設定樣本的初始樣本權重，以提升識別實際上有回堵的能力。





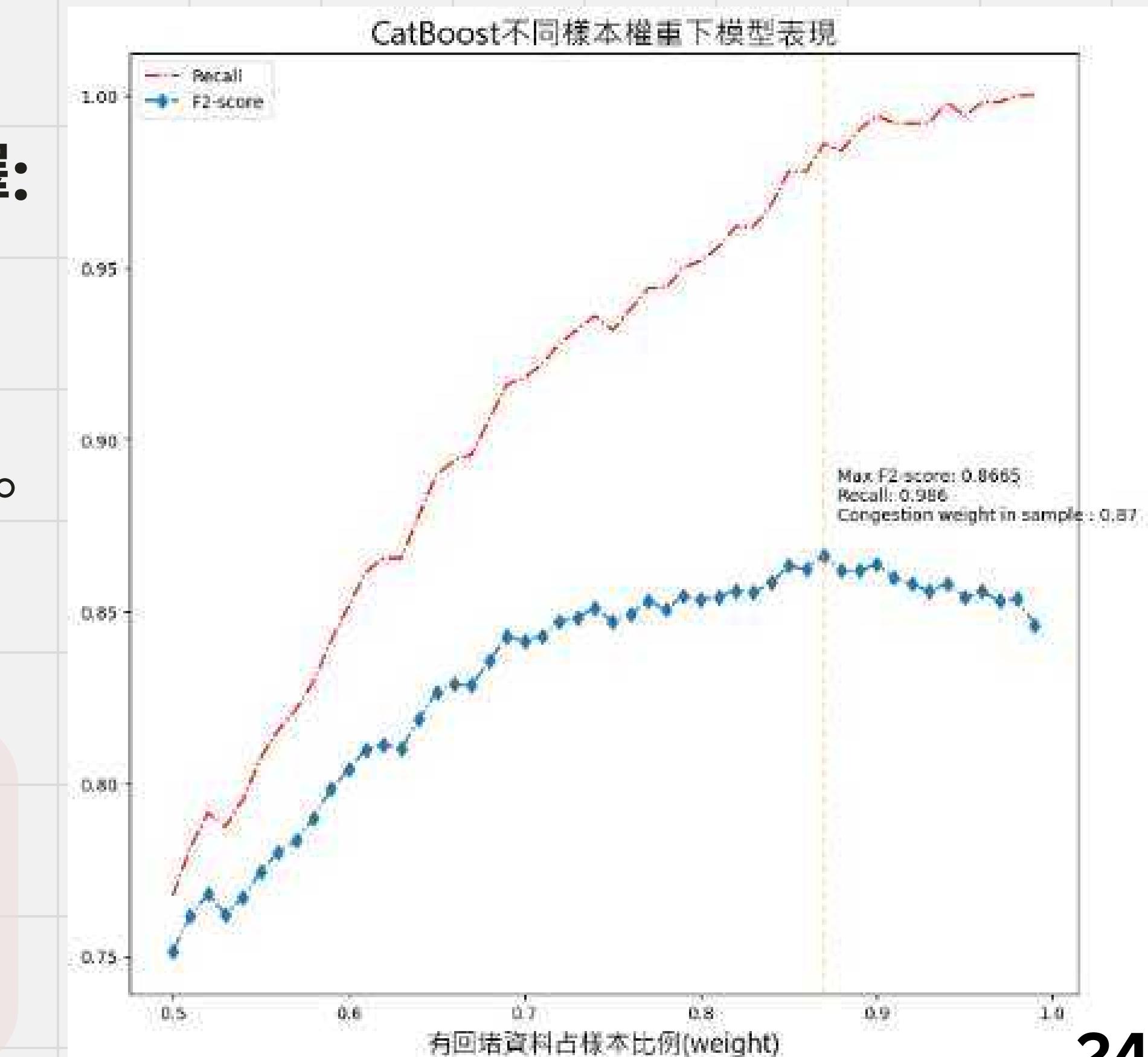
CatBoost樣本比例選擇

- 有無回堵的初始樣本比例參數選擇：

指標主要參考 **Recall** 和 **F2-score** 值

因希望真實會回堵的事件被預測回堵。

由右圖黃色虛線，當初始樣本有回堵比例為 0.87、無回堵占 0.13 時 F2-Score 最大，而 Recall 值高達 0.986 符合預期





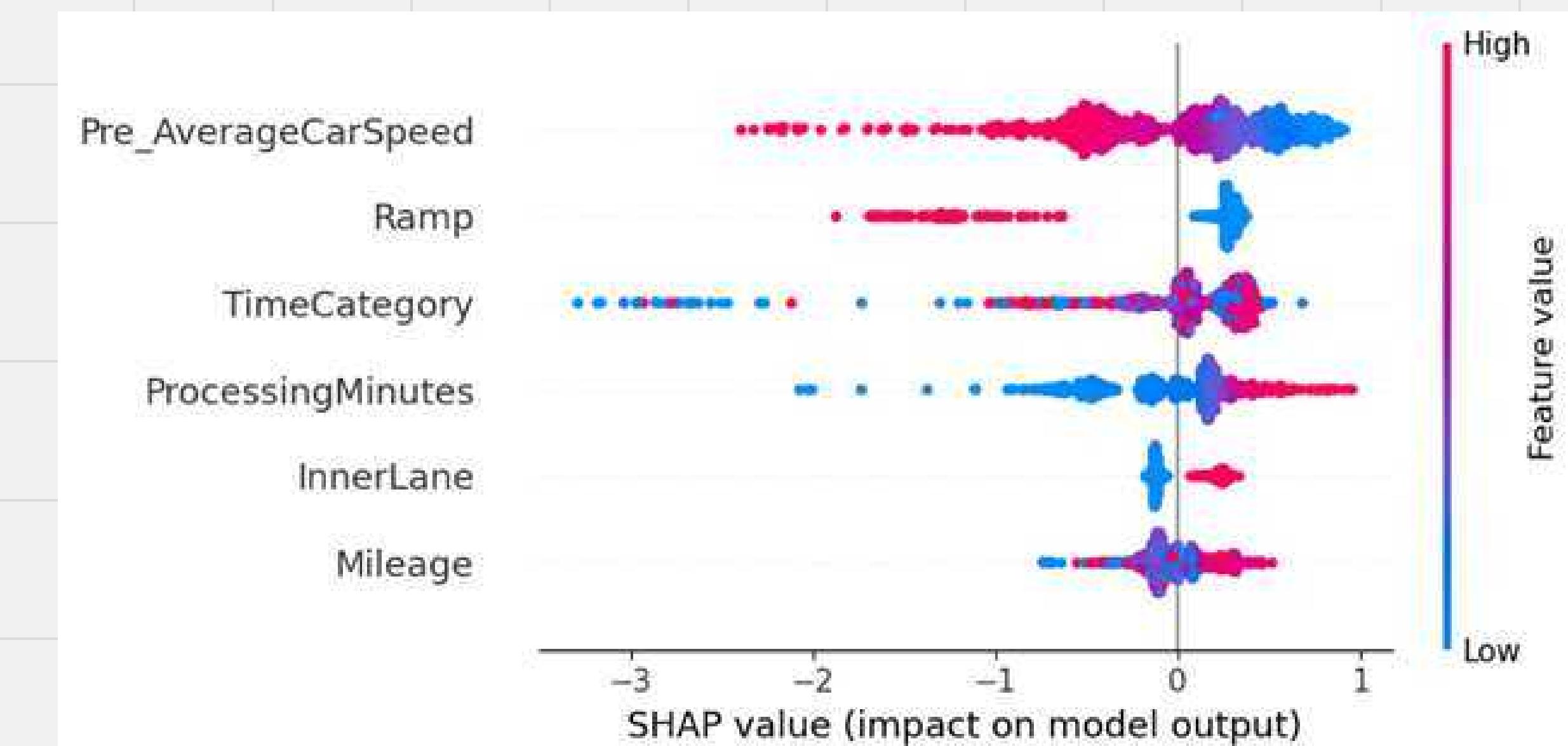
第一階段模型重要變數討論

- Shapley Value可計算不同特徵對目標問題的預測之貢獻程度

前三重要變數

!!!

- 事故前平均車速:
越低回堵越長
- 事故是否在匝道:
事故在匝道較不易回堵
- 時間類別:
特定時間對預測貢獻大





第一階段模型小結

目標

預測此事件是否會造成回堵，屬於類別型預測

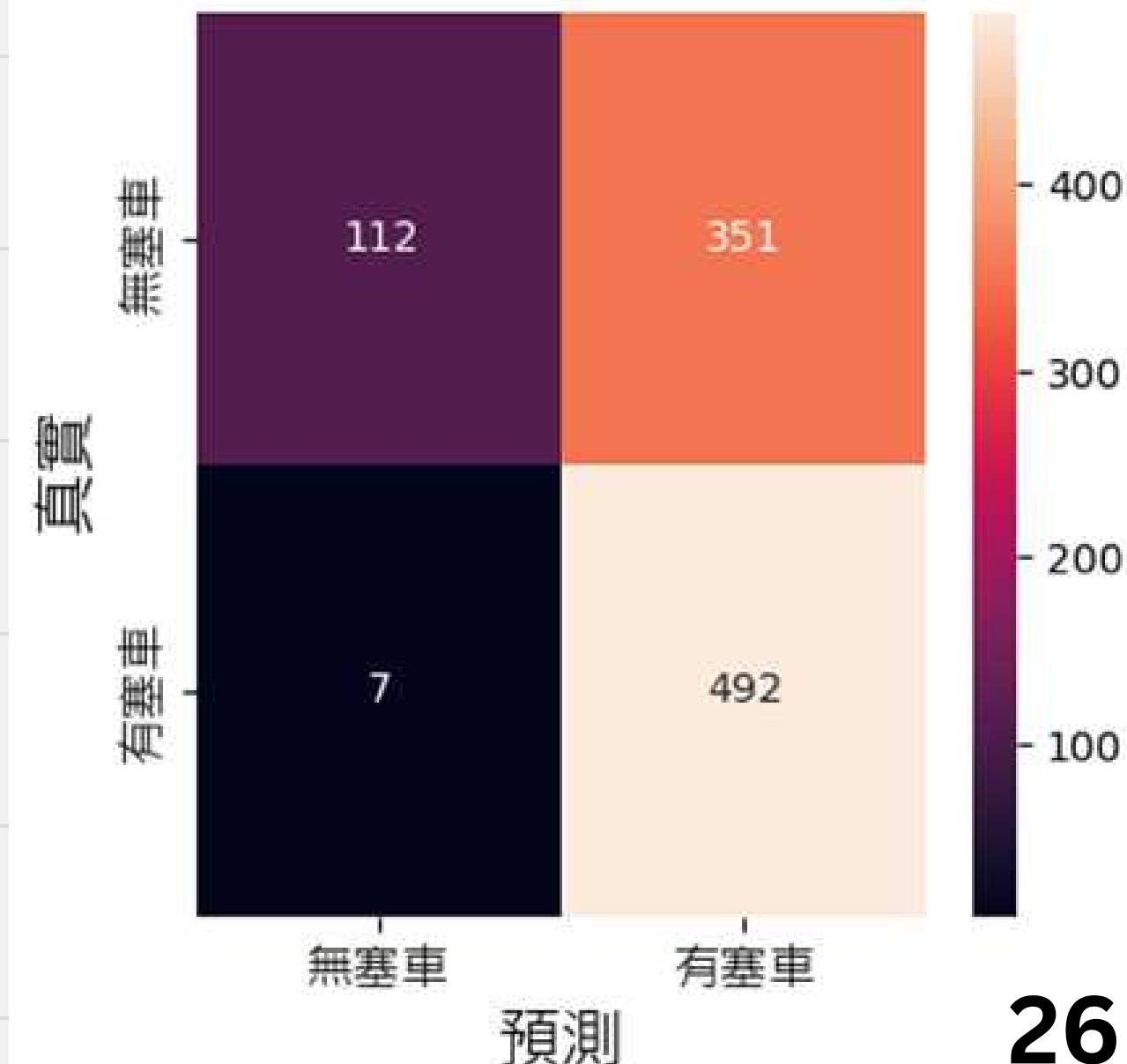
模型

- CatBoost
- 950顆樹
- 深度為4
- 初始樣本比例(有回堵: 0.87、無回堵0.13)

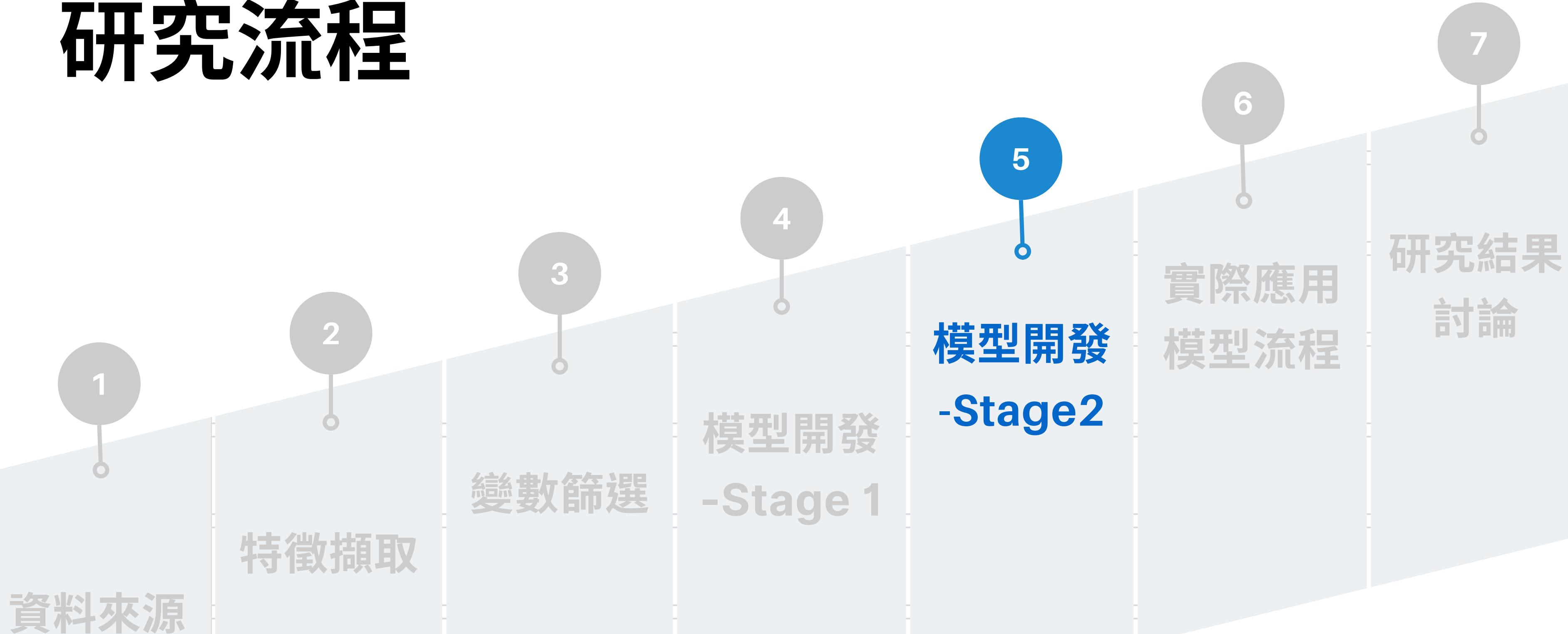
表現

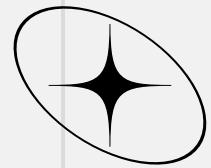
Recall 九成以上，FN處降至7筆
表會造成回堵的事件此模型高機率
預測會回堵。

CatBoost調整樣本比例後之混淆矩陣



研究流程





第二階段模型

目標

預測「Congestion Mileage」（回堵里程）

本階段使用之訓練集

112 年 1~10 月
車禍**有造成回堵的數據**

本階段使用之測試集

113 年 1~2 月 **First Stage Model** 預測會回堵之數據

第二階段模型選擇指標

- **Minimum RMSE**：降低預測的回堵里程與實際回堵里程的差距

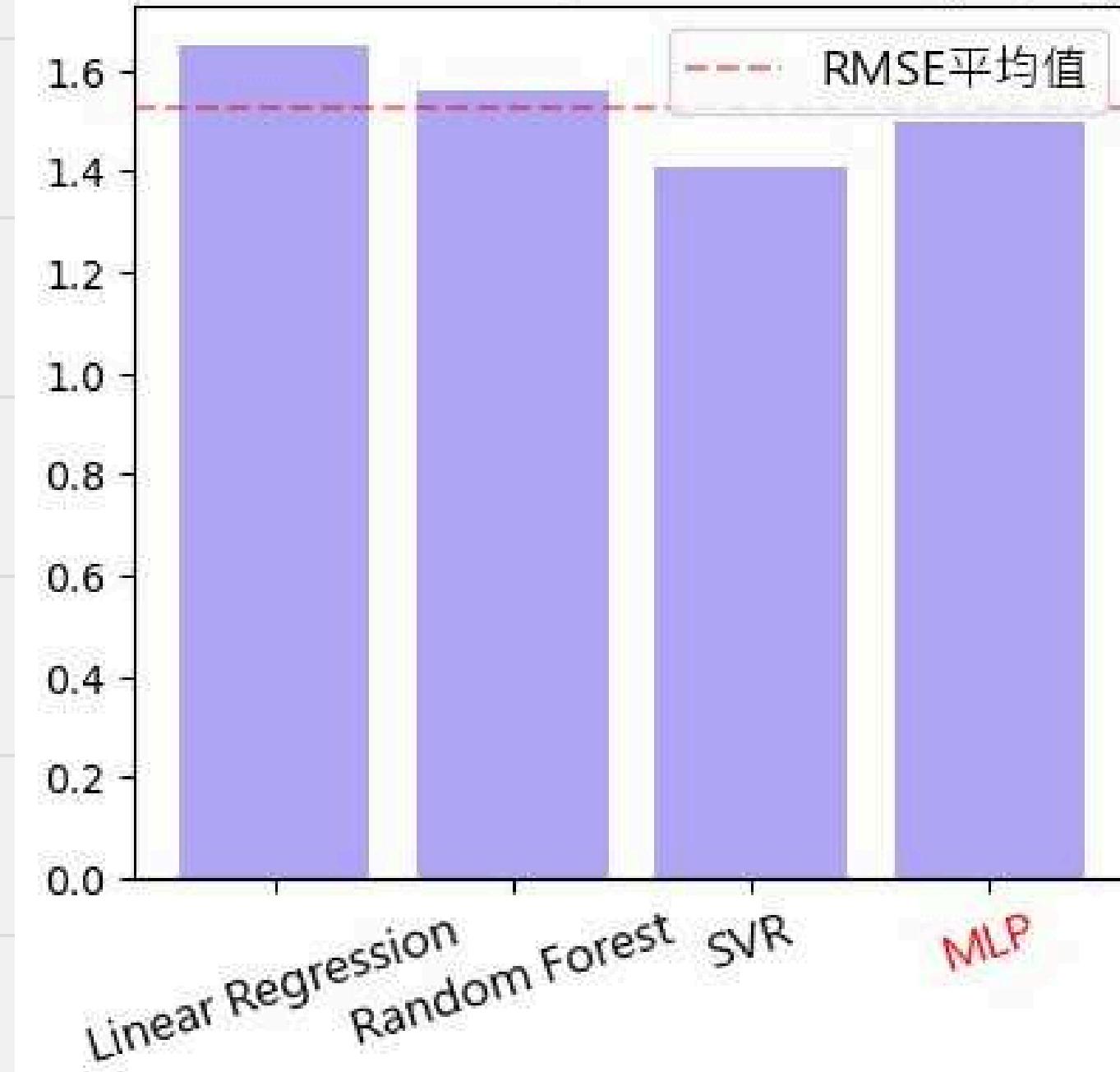
$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- **Minimum 低估比例**：降低低估回堵嚴重程度的比例

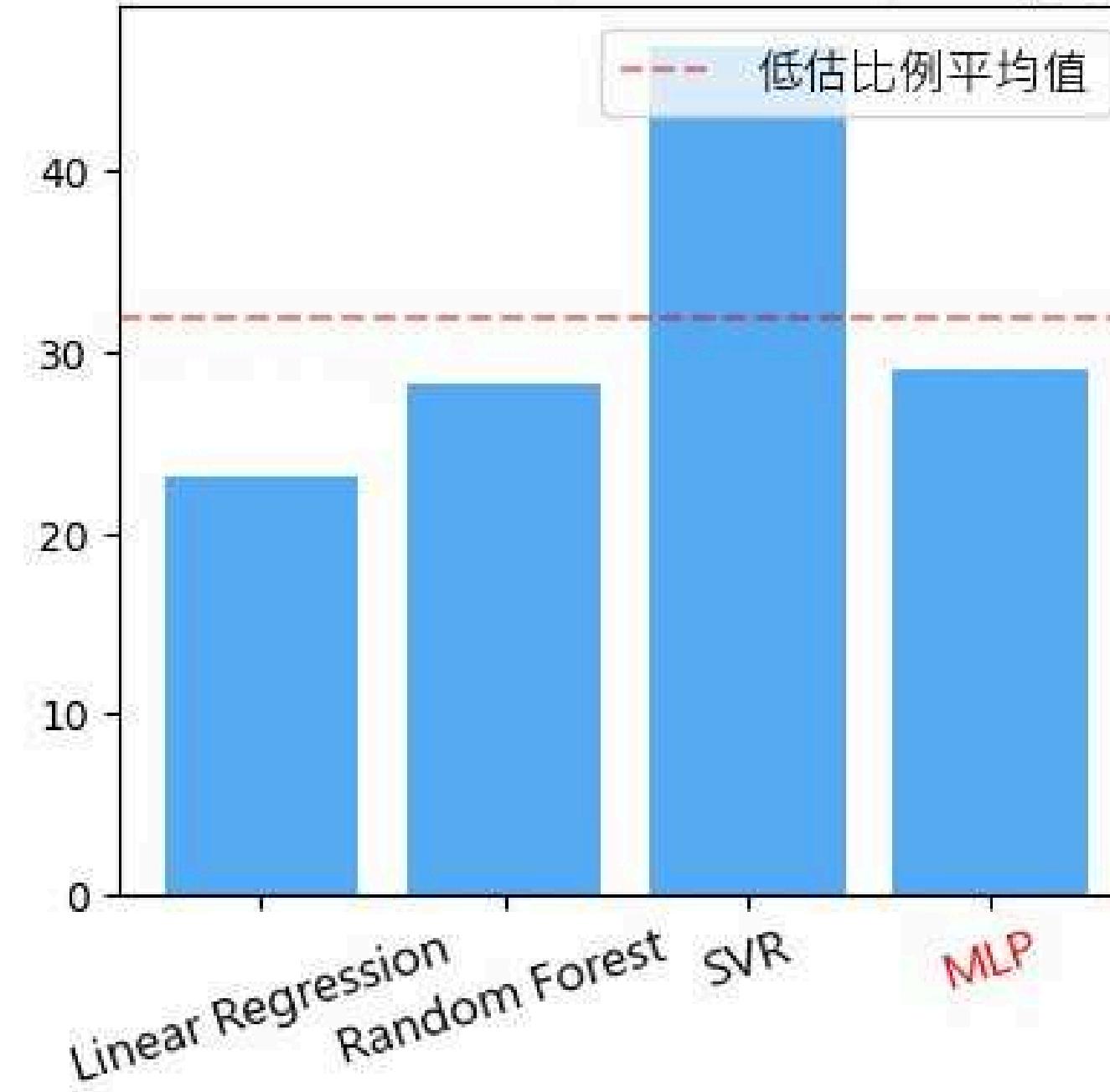
$$\text{低估比例} = \frac{\text{模型預測之「回堵里程」} < \text{實際「回堵里程」的資料總筆數}}{\text{測試集總筆數}} \times 100\%$$

第二階段模型選擇

第二階段模型比較-RMSE(公里)

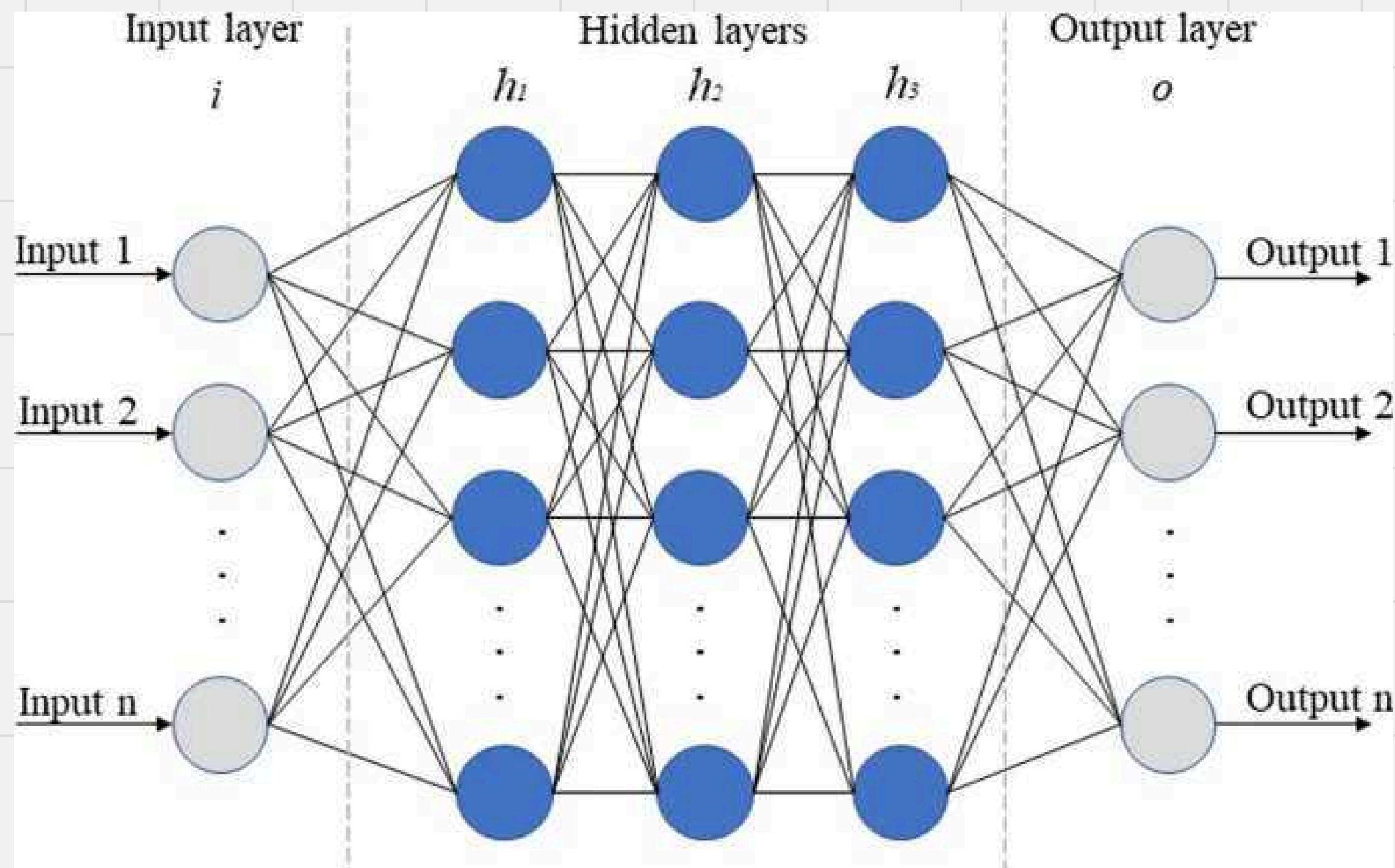


第二階段模型比較-低估比例(%)



MLP Model 在 RMSE, 低估比例的表現都優於平均

MLP模型架構



MLP模型結構圖

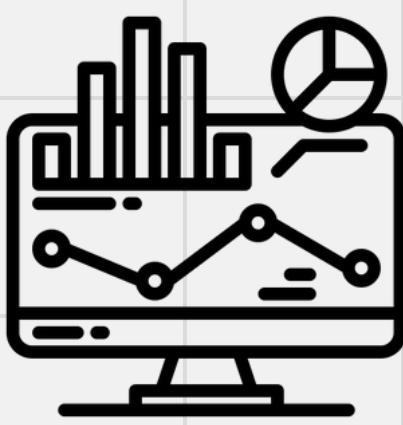
- 二階段模型結構如下：

	節點數
第1層	256個
第2層	128個
第3層	128個
第4層	64個
第5層	64個
第6層	32個

第二階段模型表現

- 對於MLP模型效果較好之原因，猜測為以下幾點

- 1.非線性建模能力：此模型能夠捕捉輸入特徵與目標變數間的複雜關係
- 2.多層特徵學習：透過多層隱藏層來學習特徵關係，效果遠超單層模型
- 3.過擬合的處理：適當於模型中加入隨機丟棄層幫助減少過擬合
- 4.Adam優化器：透過自適應學習率調整策略，更有效地找到最佳權重



第二階段模型小結

目標

預測此事件的回堵里程，屬續連續型預測

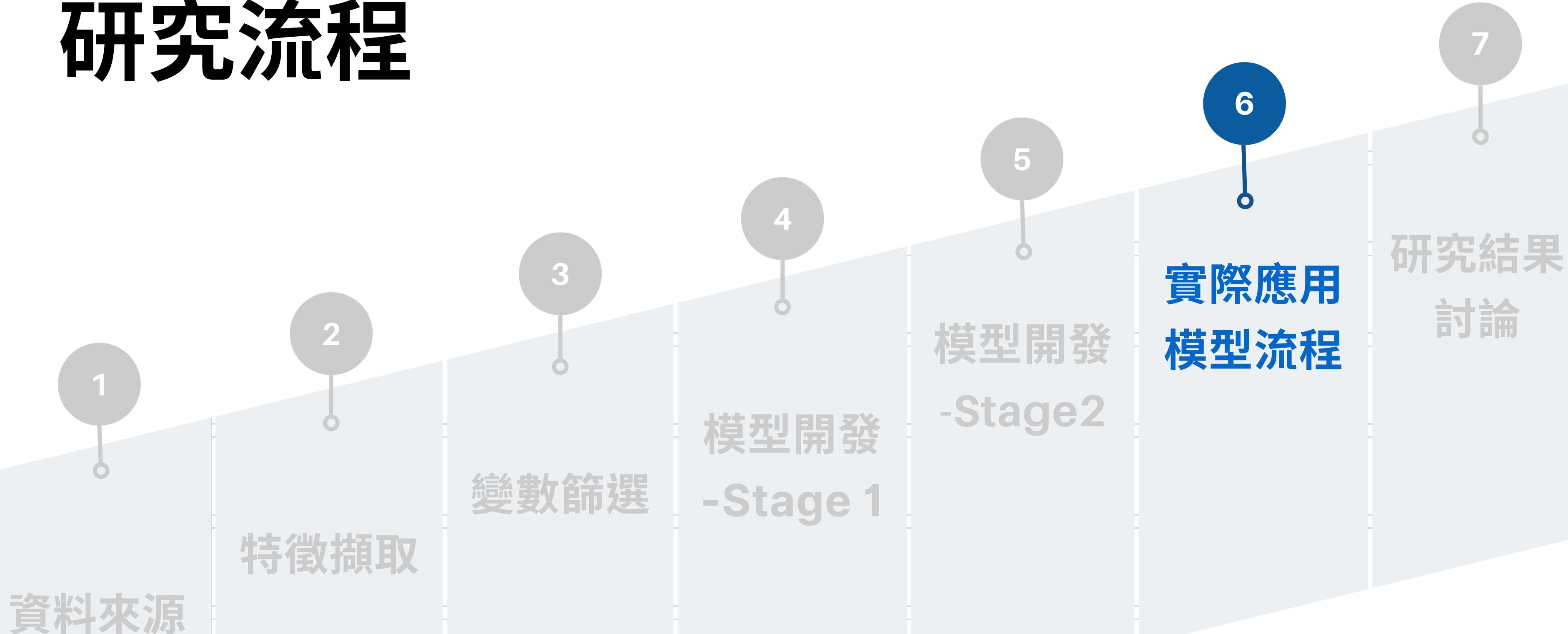
模型

- MLP
 - 150 個訓練週期、批量大小為 128 以更新模型權重
 - 輸入層、六層隱藏層（分別為 256、128、128、64、64、32 個神經元）、輸出層
 - 修正線性單元（ReLU）激活函數、Dropout 層
 - Adam 優化器

表現

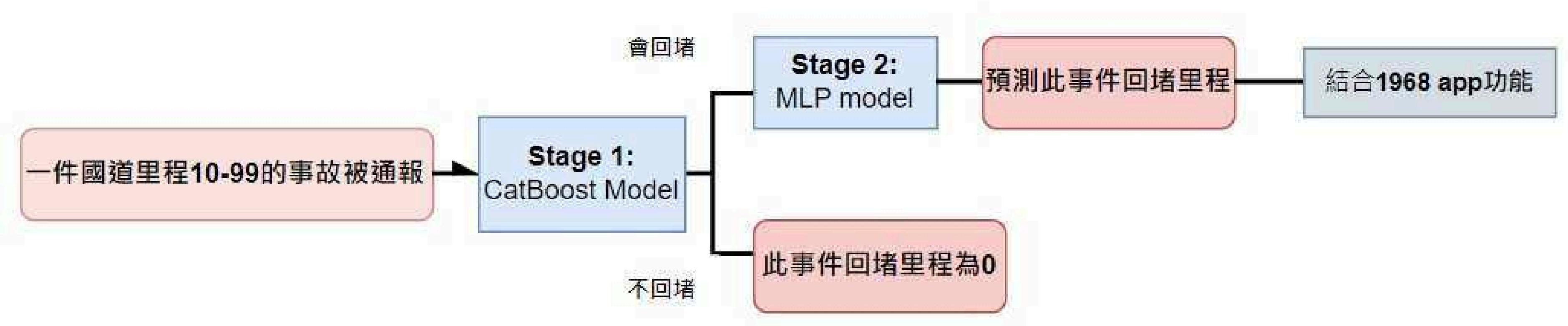
RMSE = 1.496K, 低估比例 = 29.024%

研究流程



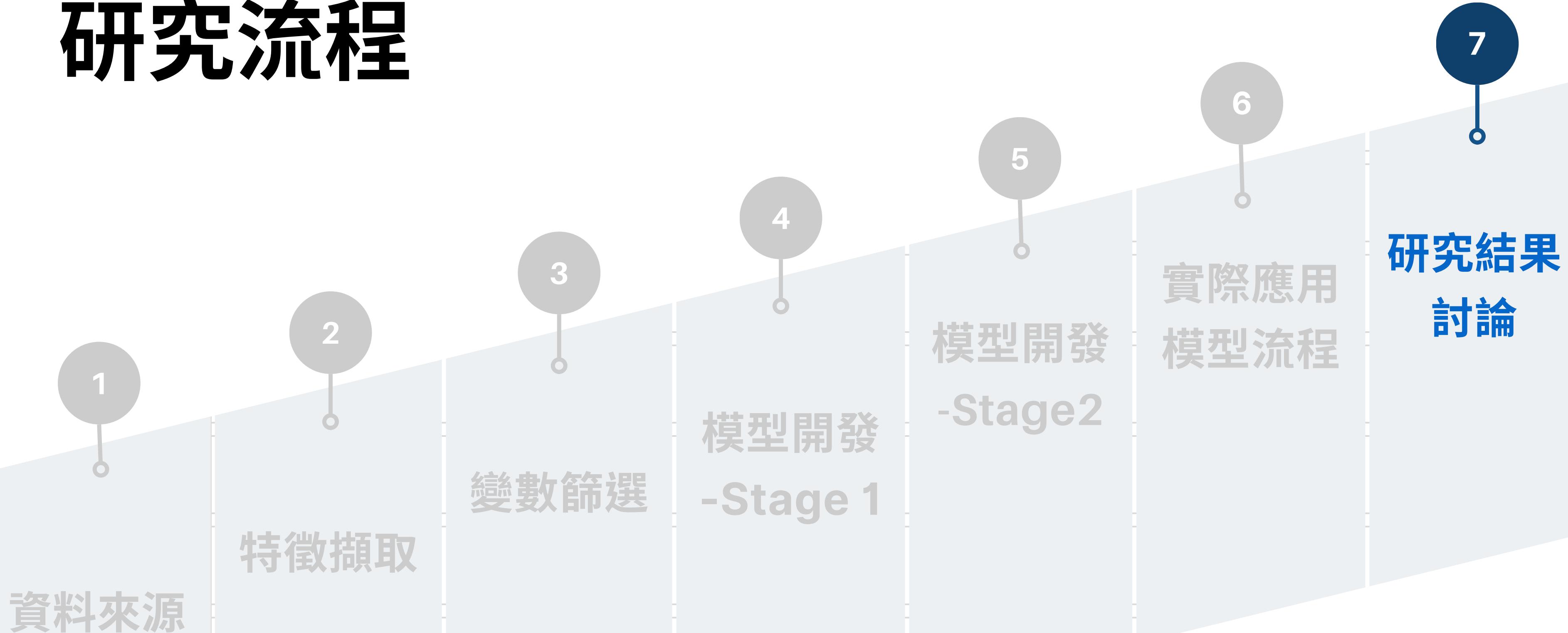


實際應用模型流程



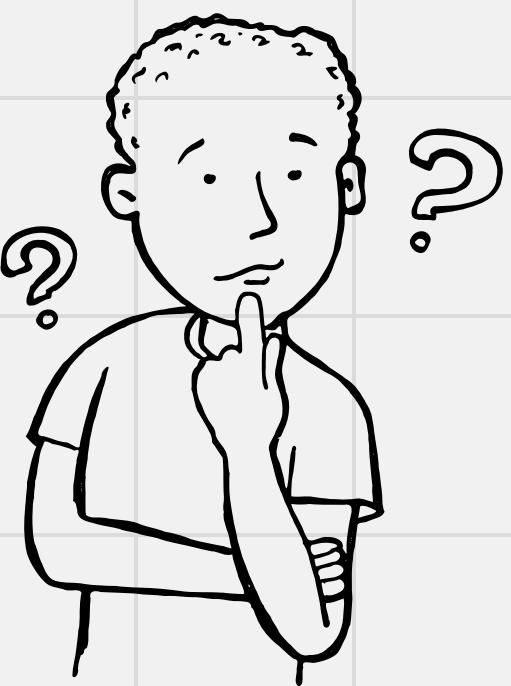
對2024年1月至2月的事件做此流程測試
RMSE約1.45K、低估比例約24.94%。

研究流程





研究結果及討論



目前研究討論

1. 測試集RMSE約1.45K，在國道上我們認為是可接受誤差，但低估比例約24.94%仍有進步空間。
2. 重要變數主要為：
 - **事故因子**：事故發生位置、時間點、處理時間、里程數
 - **事故前路況**：事故前平均車速
3. 誤差原因探討：
 - 特徵缺乏精準度：由「事故前平均車速」為貢獻最高，但氣候、車流量卻低而推論
 - 國道門道故障：資料遺失，必須提取離事故更遠的門道
 - 簡訊內容格式不一、複雜：例如回堵長度資料遺失、車種的水泥車和預拌混泥土車等...



研究結果及討論

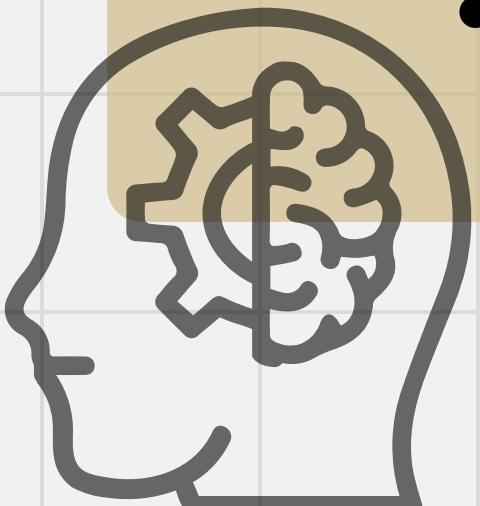
未來展望

1.降低低估比例：

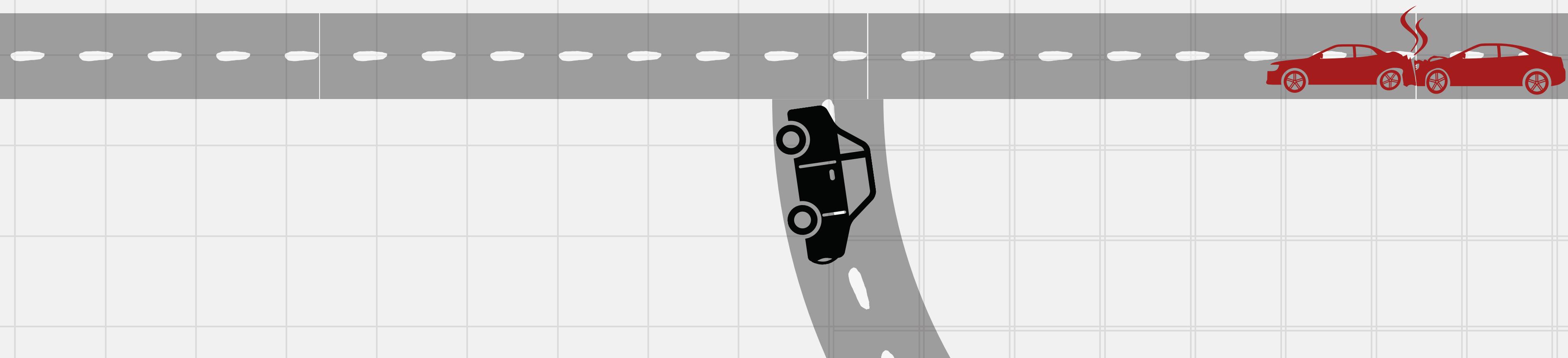
- **第二階段模型**：因第一階段的Recall已經夠高，我們認為可以對第二階段建模改良，例如：把低估比例及MSE加入機器學習的loss function中等.... 試圖降低低估比例。

2.降低預測誤差：

- **特徵方面**：針對預測事故蒐集相關數據，如：每分鐘車流量、每分鐘車速、每分鐘降雨量等....，以及對門道設計損壞備案。



實務應用與可行性



即時路況與路網圖



當前功能：
以一分鐘的頻率更新當前高速公路各路段時速，並以顏色區分不同平均時速。



可添加功能：
利用高頻率更新的優勢，於事故發生**當下**提供位置標誌，並附上**事故詳情**，包含**回堵里程**予以警示。

路段時間預測



當前功能：
可以依照所選定的交流道計算旅途時間，或是在訂閱特定路段後定期收到時間預估通知。



可添加功能：
於計算過程中增加事故回堵因素，並列出有無事故下的行車時間差異，減少低估時間的可能性。

路況事件推播



當前功能：

會定位目前車輛位置，
對前方特定公里內發生
的事件，如管制、施工
等進行語音推播。

NEW

可添加功能：

在事故的**推播**內容中可
加入回堵公里資訊，以
及遭遇回堵前的交流
道，提供駕駛迴避壅塞
路段的交流道選項。

113年國道智慧交通管理創意競賽



THANK YOU

