題目: 國道 1 號里程 10~99k 的車禍回堵里程預測

隊伍名稱：我的獎金~~15萬！

目錄

摘要

本研究使用機器學習技術對事故發生後的回堵里程預測，以降低預測誤差及低估比例為目的，不僅使用了競賽主辦方提供的資料，更結合高公局資料庫中的交通即時資訊和外部天氣資料庫的天氣概況，增加了變數的複雜程度，也提供模型更好的即時因子，使預測結果可以更加符合實際情況。

基於CatBoost、多層感知器等模型架構，發展出二階段回堵里程預測模型，研究流程可參考圖1，根據氣候、交通即時、事故、施工等資訊，首先判斷此事故是否即將發生道路雍塞，而對於預測會回堵的事故，再進一步預測路段回堵里程長度，最終測試集結果RMSE為1.45公里，而低估比例約24.94%。

最後依本研究成果性質，建議和高公局開發的 1968 App 結合，為各個功能添加即時事故回堵的影響資訊，利用更新率快的優勢提早通知未來可能發生的交通壅塞的路段，以及利用路人位置的推播功能，對即將駛向回堵區域的駕駛發送語音提醒，以提供替代路線建議等，緩解大台北地區與鄰近區域的交通有幫助。

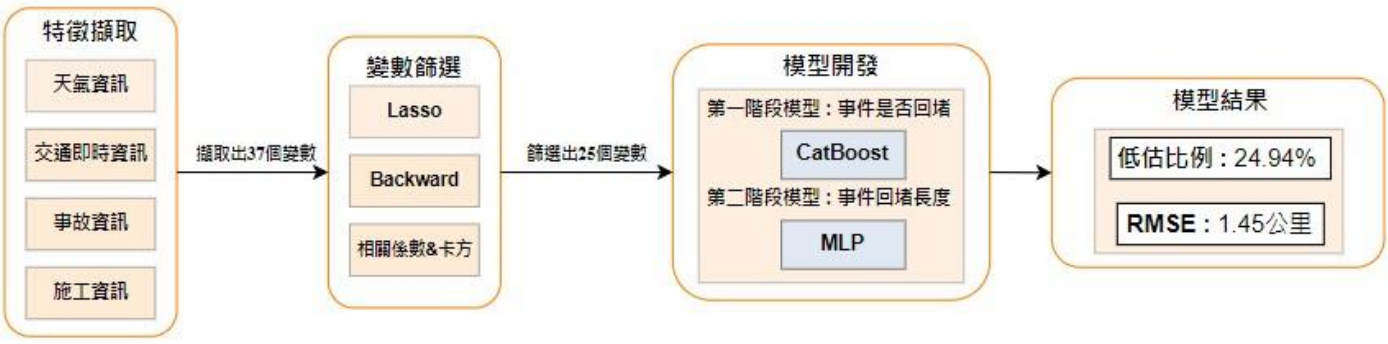


圖 1 研究流程圖

1引言

台灣的國道系統涵蓋全台主要城市和重要經濟區域，提供快速便捷的交通服務。然而隨著車輛數量的增加和經濟活動的繁忙，國道系統將面臨日益嚴重的壅塞問題。

台灣的國道總長度約為 1,000 公里，涵蓋多條主要高速公路，包括國道一號、國道三號、國道五號等。這些國道連接了北、中、南三大經濟圈，每天承載著數百萬輛的車流量，其中又以國道一號為全台最重要的南北向運輸幹道。國道的交通壅塞主要發生在上下班高峰時段、週末及假期，尤其是在主要都市和工業區附近路段，根據交通部的數據，國道一號途經許多人口密集區，如台北內湖、新北林口、桃園南崁和新竹工業園區等，經常出現嚴重的交通壅塞現象。

基於交通部公告之易壅塞路段彙整表[1]，本次研究**主要關注國道一號的里程 10~99 公里路段，同時涵蓋了上述的人口密集區**以及**易壅塞路段**。這段路段的交通壅塞主要由以下幾個原因造成：

1. **高車流量**：上下班高峰時段的車流量急劇增加，遠超過路段的設計容量。
2. **頻繁的進出口**：該路段設有多個進出口，車輛在進出時頻繁變換車道，容易造成交通瓶頸。
3. **事故頻發**：高流量和複雜的交通環境使得該路段事故多發，而每次事故都會進一步加劇交通壅塞。

本研究將針對國道一號里程 10~99 公里路段的交通事故造成的回堵里程進行預測，期望通過精確的數據分析和模型建構，提供有助於減緩壅塞的有效措施和建議。

2 文獻回顧

1. A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data[2]：

此文獻將車流量壅塞程度分為五個等級，並以警察抵達車禍處的時間點切分成二階段建模，研究目的為事故後需要多少時間才能回到事故前的壅塞等級。

特徵主要以交通即時因子、事件因子、天氣因子為主，第一階段建模因警察還未

抵達，故只有交通即時因子及天氣因子可供建模，並以二分法（Binary）及多元分類（Multiclass classification）為主，預測壅塞程度是否會更加嚴重，本研究的測試集準確率隨道路原本的壅塞程度不同在 0.73 至 0.83 之間。**第一階段重要因子多為交通即時因素**，分別為車禍發生時前後一分鐘車速差、事件發生前五分鐘到事件發生時的車流量、事件發生當下的降雨量以及事件發生前五分鐘到車禍發生的大型車比例。

若第一階段預測壅塞程度是會更加嚴重，則在警察調查完畢後進入第二階段模型，主要預測仍是需要多少時間回到原本的壅塞程度，測試集的 AD (Absolute difference)大多低於 10 分鐘，**第二階段重要因子主要集中於事故因子**，如是否有人受傷、事故原因是否為追撞、警察抵達後一分鐘的平均時速、事故車總數、天氣概述。

此文獻**以二階段先預測類別再討論連續型是很好的想法，但在最後第二階段處理時間的誤差上沒有討論高估與低估**，且第一階段的準確率仍有進步空間。

2. Congestion Prediction With Big Data for Real-Time Highway Traffic[3] :

此文獻針對即時高速公路交通擁堵預測進行了深入的研究，在特徵因子的選擇上，研究考慮了多個關鍵因素，包括路速、路段密度、交通流量、降雨量以及即時的交通事件報告。而我們的研究也參考該文獻加入這些特徵因子，不僅能反映當前的交通狀況，還能幫助模型更準確地預測未來的交通情況。

此文獻模型選擇使用支持向量機（SVM）建立即時高速公路交通擁堵預測模型（SRHTCP）。研究結果顯示 SRHTCP 模型在預測準確性上顯著優於基於加權指數移動平均法的預測方法，具體表現在預測準確性提升了 25.6%。在平均絕對相對誤差（MARE）的測量中，該模型表現相當不錯。SRHTCP 模型不僅能夠即時預測下一時間段的車速，還能有效分析高速公路的交通擁堵情況。通過整合交通、天氣和社交媒體數據，該模型能夠全面地反映交通狀況並提升預測的可靠性。

3 研究方法及步驟

3.1 資料來源

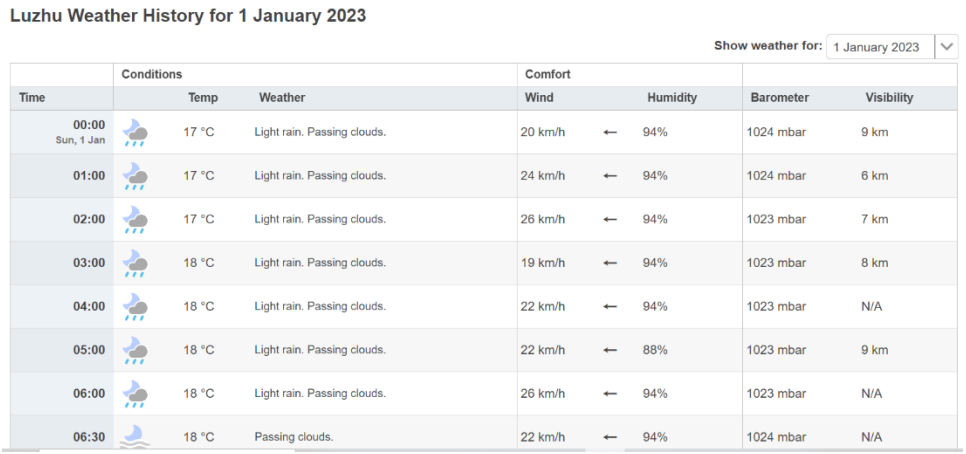
本研究使用了 Timeanddate 網站[4]中的天氣因子，包含溫度(℃)、風速(km/h)、濕度(%)以及氣壓值(mbar)；交通部高速公路局交通資料庫[5]中的 Etag 靜態資訊(v2.0)、M03A(各類車種通行量統計)、M06A(各旅次路徑原始資料)；112 年 1-10 月及 113 年 1-

2 月道路施工路段資料[6]；112 年 1-10 月及 113 年 1-2 月國道 A1、A2、A3 交通事故資料；112 年 1-10 月及 113 年 1-2 月交通事故簡訊通報狀況資料[6]。**透過國道一號及 10~99k 的里程篩選後，**擷取出了 **5702 筆的事故事件做為訓練集**和 **962 筆事故事件做為測試集**。

3.2 特徵擷取

3.2.1 **天氣資料** ：

Timeanddate 網站有對台灣各地區、每小時的詳細氣象資訊（圖 2）。



Luzhu Weather History for 1 January 2023

Show weather for: 1 January 2023

	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

圖 2 Timeanddate 網站資料呈現

利用網路爬蟲針對每起事故提取重要的天氣因子，包含**溫度(℃)、風速(km/h)、濕度(%)以及氣壓值(mbar)**。

另外，由於網站的氣象資料是由地區區分，也沒有關於氣象測站的詳細資訊(如經緯度)，故本研究只能由里程分類至可查詢地區，如下表。

國道一號里程	地區	國道一號里程	地區
10 - 15	汐止區	49 - 52	大園區
15 - 17	內湖區	52 - 62	中壢
17 - 25	中山區	62 - 71	平鎮區
25 - 27	三重區	71 - 86	湖口區
27 - 35	五股區	86 - 91	竹北區
35 - 41	林口區	91 - 95	新竹東區
41 - 49	蘆竹區	95 - 99	新竹縣寶山

表 1 國道 1 號里程與地區對應表

3.2.2 **國道車流量及平均車速特徵擷取** ：

利用高公局資料庫的 Etag 靜態資訊(v2.0)[4]，並鎖定國道一號的測站。另外，為避免測站失靈故多提取里程數 10-99 外的測站，資料樣式如下圖 3：

	測站代碼	國道	方向	路段起點描述	路段迄點描述	里程數
0	01F0061S	國道1號	S	大華系統	五堵	6.1
1	01F0061N	國道1號	N	五堵	大華系統	6.1
2	01F0099S	國道1號	S	五堵	汐止&汐止系統	9.9
3	01F0099N	國道1號	N	汐止&汐止系統	五堵	9.9
4	01F0147N	國道1號	N	東湖	汐止&汐止系統	14.7
...	...	...	...	...	...	...
66	01F0980S	國道1號	S	新竹(科學園區)	新竹系統	98.0
67	01F1045N	國道1號	N	頭份	新竹系統	104.5
68	01F1045S	國道1號	S	新竹系統	頭份	104.5
69	01F1123N	國道1號	N	頭屋	頭份	112.3
70	01F1123S	國道1號	S	頭份	頭屋	112.3

圖 3 資料樣式示意圖

測站定義（示意圖如圖 4）：

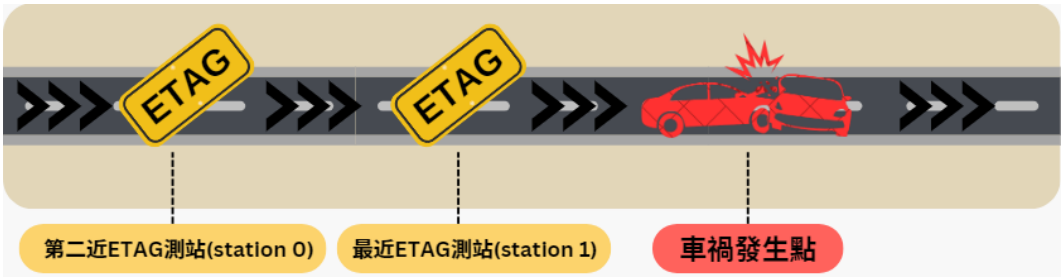


圖 4 測站定義位置示意圖

由圖 3 可定義兩種測站分別為：

- station 0: 車子已經過且離事故第二近的測站
- station 1: 車子已經過且離事故最近的測站

以利後續特徵擷取解釋。

(1) **事故前十分鐘的車流量**(Pre_TrafficVolume)及**大型車比例**(Pre_LargeVehicleRatio)：

此**特徵意義為評估事故發生前道路的路況**，主要從高公局 M03A 提取，此資料（圖 5）是以每五分鐘為時間區段，即檔案中時間區段為 2023-08-17 06:00 代表 2023-08-17 06:00 – 06:05 的總車流。

	時間區段	門架代號	方向	車種	車流量
0	2023-08-17 06:00	01F0005N	N	小客車	19
1	2023-08-17 06:00	01F0005N	N	小貨車	7
2	2023-08-17 06:00	01F0005N	N	大客車	3
3	2023-08-17 06:00	01F0005N	N	大貨車	0
4	2023-08-17 06:00	01F0005N	N	聯結車	0

圖 5 每五分鐘為時間區段的總車流

本研究針對該事件找出其 station 1 後，找出該門架的事故時間前十分鐘的車流量（如式一、式二）資料（不包含事故發生當下的時間區段）。如:

- 事故發生在 03:04 分提取 02:50 及 02:55 資料
- 事故發生在 03:05 分提取 02:55 及 03:00 資料
- 事故發生在 03:00 分提取 02:50 及 02:55 資料

$$\text{總車流量} = \text{小客車車流量} + \text{小貨車車流量} + \text{大客車車流量} + \text{大貨車車流量} + \text{聯結車車流量} \quad \text{（式一）}$$

$$\text{大型車比例} = \frac{\text{大客車車流量+大貨車車流量+聯結車車流量}}{\text{總車流量}} \quad \text{（式二）}$$

(2) **事故後兩分鐘的車流量**：

此**特徵意義為計算事故發生後會有多少車再進入事故區**，對於預估塞車長度可能會有幫助，由於車流量可使用資料有限，故主要從高公局 M03A 提取，但因該時間區段為五分鐘，本研究採取權重平均擷取，如 :

- 事故發生在 03:32 分提取 03:30 – 03:35 車流量資料並取 2/5。
- 事故發生在 03:29 分提取 03:25 – 03:30 及 03:30 – 03:35 車流量資料並各取 1/5。
- 事故發生在 03:35 分提取 03:35 – 03:40 車流量資料的 2/5。

(3) **事故前十分鐘平均車速**(Pre_AverageCarSpeed)：

此**特徵意義為評估事故發生前路況**，提取自高公局 M06A，該資料由門架追蹤車輛，並記錄經過測站的時間點，並以每小時為時間區段，即檔案中的時間區段為 2023-12-23 04:00 代表在 2023-12-23 04:00:00 – 04:59:59 經過起始門架的車輛資料。

記錄方式如'2023-12-23 04:34:39+01F0928N; 2023-12-23 04:37:38+01F0880N; 2023-12-23 04:45:52+01F0750N; 2023-12-23 04:54:54+01H0608N; 2023-12-23 04:56:25+01F0584N'，代表該車在 2023-12-23 04:34:39 經過代碼為 01F0928N 的測站(國道一號 里程 92.8)，2023-12-23 04:37:38 經過代碼為 01F0880N 的測站(國道一號 里程 88.0)，以此類推。

提取單一事件發生前十分鐘經過 station 1 的總行車資料，共 n 筆，由測站里程及時間差計算單一車在事件發生前車速，如式三。

$$speed_j = \frac{\text{測站里程差(km)}}{\text{時間差(s)}}, j = 1 \ldots\ldots n \qquad \text{（式 三）}$$

由上例的資料，若要提取的 station 1 為 01F0928N 且 04:34:39 在事故發生前十分鐘內則記錄其行駛車速， 即(92.8 - 88)/(175) = 0.0242(km/s)。

由式四，再對 n 筆車輛行駛速度取平均代表單一事件前十分鐘的行駛車速

$$\text{Pre\_AverageCarSpeed}_i = \frac{\sum_{j=0}^{n} speed_j}{\text{n}}, i = 1 \ldots\ldots 5702 \qquad \text{（式四）}$$

然而，雖然 M06A 能協助提取時間較精準的速度資料，但仍有些挑戰。下述為處理 M06A 所遇到的問題：

1. 鎖定小客車車種 ：

因單一檔案資料量大，每小時內容量至少十萬筆，但由於目的是計算平均速度，且國道上大部分車種為小客車，較不易有資料缺失問題還能先快速篩選資料，故此特徵本研究針對”小客車”的車種計算速度。

2. 測站失靈 ：

由於 ETAG 門架有時因故障失靈，導致該時段 station 1 資料遺失，若遇到此情形本研究往前一個測站，即 station 0 計算特徵。

3. 單一事件計算資料量大 ：

因 M06A 區分資料的方式為“該車經過起始門架的時間點” ，故檔案若只鎖定事件那一小時的檔案是不夠的，且有可能會沒有資料，舉例：有可能車子最初行經該起始門架的時間點為 12:00，可是 15:00 才經過我們的目標門架，故本研究往前提取四小時內的檔案計算。例如 ： 事件發生在 04:16 分，會抓取 12:00:00、01:00:00、02:00:00、03:00:00、04:00:00 的 M06A 車輛行駛資料。

3.2.3 事故簡訊資料

資料是來自國道智慧交通管理創意競賽網站[6]所提供的 112 年 1-10 月交通事故簡訊通報資料.xlsx 與 113 年 1-2 月交通事故簡訊通報資料.xlsx。

本研究先將重複紀錄的資料清除，因為資料表內紀錄的部分內容與其中的”簡訊內容”欄位紀錄的不同，因此決定統一採用”簡訊內容”中的紀錄當作資料，接著依照關鍵字拆分簡訊，再逐項處理。

(1) 時間類別 (TimeCategory)、日期類別 (DateCategory)、處理分鐘 (ProcessingMinutes) :

依照簡訊擷取事件發生的年、月、日、時、分、和事件發生及排除的時間，將事件發生小時儲存成時間類別，而日期類別則是依日期劃分平日、一般週六日及國定假日，最後再計算事件排除所需時間紀錄成處理分鐘。

(2) 方向 (Direction) :

依照簡訊擷取出北、南、東、西及雙向五種，因為本次研究只使用國道一號部分路段，因此方向變數只有包含南向與北向兩種。

(3) 里程 (Mileage)、**回堵里程** (CongestionMileage) :

依照簡訊擷取車禍事件發生的里程位置，再根據有無回堵的結果，**無回堵表示為 0，而有回堵則另擷取其回堵里程**，單位皆為 km。

(4) 死亡 (Deaths)、受傷 (Injuries):

依照簡訊擷取該車禍事件是否有人員傷亡，分別記錄死亡和受傷人數。

(5) 事故波及車道資料 (InnerShoulder, InnerLane, InnerMiddleLane, MiddleLane, OuterMiddleLane, OuterLane, OuterShoulder, Ramp) :

因為簡訊內並無詳細記錄事故波及車道，因此使用資料原本的八個車道欄位。

(6) **車輛相關欄位** (AccidentVehicle, ConstructionVehicle, Car, MediumLargeBus, MediumSmallTruck, LargeTruck_Trailer, OtherVehicles) :

透過擷取每條簡訊中的肇事以及受波及的車輛名詞，將每輛車獨立紀錄，因為每條簡訊的紀錄名詞都不同，同為小客車的欄位就包含如小客、小客車、小車、小自客…等，因此將同一類車輛名詞統一後，**分類成工程用車輛、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**，再按每筆事件資料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。

(7) **事故相關欄位** (RearEndCollision, SelfCollision, Fire, Overturn, OtherCause) :

透過擷取每條簡訊中的肇事原因，分類為追撞、自撞、擦撞、打滑、翻車、散落物、起火、其他等八類，因為有少數的事件擷取出的肇事原因同時包含兩項以上，如簡訊內容同時提及散落物和自撞，因此再合併類別為**追撞、自撞、翻車、起火、其他等五項**。

3.2.4 施工資料

資料是來自國道智慧交通管理創意競賽網站[6]所提供的 112 年 1-10 月道路施工路段資料.xlsx 與 113 年 1-2 月道路施工路段資料.xlsx。

先透過國道及里程篩選出與此次選定路段相符的施工資料，並分成南向與北向車道，分別記錄各施工工程占用的車道數量 (LaneOccupancy)，包含各車道、路肩和分隔島等，最後紀錄工程開始和完成時間。

至於將**施工資料與事故資料合併的標準**，篩選事**故地點的里程落在施工地點里程前後 300 公尺內，且最晚事故時間前 20 分鐘施工都還在進行**，那就認定此次事故發生的路段可能包含工程因素影響壅塞，紀錄其占用車道總數作為施工佔用車道欄位。

經過以上的資料預處理，本研究的資料欄位介紹如下表，並附上各變數於 **Lasso 算法、Backward selection、卡方檢定與相關係數**三個特徵篩選下的結果（表 2）:

欄位名稱	欄位定義	變數篩選檢定結果		
		Lasso	Backward	卡方& 相關係數
事故概述				
日期類別（DateCategory）	事故的日期類別	V	V	V
時間類別（TimeCategory）	事故的時間類別	V	V	V
方向（Direction）	事故的車道方向	V	V	V
里程（Mileage）	事故距國道起始點的里程數	V	V	V
處理分鐘（ProcessingMinutes）	從事故發生到排解的分鐘數	V	V	V
死亡（Deaths）	事故的死亡人數			V
受傷（Injuries）	事故的受傷人數	V	V	V
回堵里程（CongestionMileage）	事故排解後堵塞的里程數	此變數為目標變數，因此不加入變數篩選。		
交通即時資訊				
事故前大型車比例（Pre_LargeVehicleRatio）	事故前十分鐘大客車、大貨車、聯結車的佔比	V	V	V
事故前平均車速	事故前十分鐘的平均車速	V	V	V

（Pre_AverageCarSpeed）				
事故前車流量 （Pre_TrafficVolume）	事故前十分鐘的總車流量	V	V	V
事故後車流量 （Post_TrafficVolume）	事故後兩分鐘的總車流量		V	V
施工狀況				
施工佔用車道（LaneOccupancy）	施工佔用的車道總數			V
事故波及車道狀況				
內路肩（InnerShoulder）	事故是否波及到內路肩			V
內車道（InnerLane）	該事故是否波及內車道	V	V	V
中內車道（InnerMiddleLane）	事故是否波及到中內車道	V	V	V
中車道（MiddleLane）	事故是否波及到中車道	V	V	V
中外車道（OuterMiddleLane）	事故是否波及到中外車道	V	V	V
外車道（OuterLane）	事故是否波及到外車道		V	
外路肩（OuterShoulder）	事故是否波及到外路肩			V
匝道（Ramp）	事故是否波及到匝道	V	V	V
事故波及車輛種類（以車輛重量、車輛大小為分類依據）				
肇事車輛（AccidentVehicle）	事故波及的車輛總數	V		V
工程用車輛 （ConstructionVehicle）	事故波及的工程用車輛數			V
小客車（Car）	事故波及的小客車輛數		V	V
中大型客車（MediumLargeBus）	事故波及的中大型客車輛數			V
中小型貨車 （MediumSmallTruck）	事故波及的中小型貨車輛數	V	V	V
大貨車及聯結車 （LargeTruck_Trailer）	事故波及的大貨車、聯結車輛數	V	V	V
其他車輛（OtherVehicles）	事故波及的其他車種車輛數			
事故原因				
追撞（RearEndCollision）	事故原因是否有包含追撞			V
自撞（SelfCollision）	事故原因是否有包含自撞			
起火（Fire）	事故原因是否有包含起火			V
翻覆（Overturn）	事故原因是否有包含翻覆	V	V	V
其他事故（OtherCause）	事故原因包含其他原因			
天氣概況（以距離事故發生地理位置最近之天氣測站為主）				
溫度（Temp）	事故當下溫度			V
風速（WindSpeed）	事故當下風速		V	V
濕度（Humidity）	事故當下相對濕度			V
氣壓計（atm）	事故當下氣壓			V

表 2 資料欄位統整表

3.3 變數篩選

根據變數檢定結果，做以下變數調整：

(1) **傷亡** (Casualties) :

受傷的檢定結果不差，而死亡則相當差，考量到有死亡的事故筆數相當少，可能影響到檢定結果，因此**合併死亡及受傷**，新增傷亡人數變數。

(2) **肇事車輛數** (AccidentVehicle)、大型車數量 (LargeVehicle) :

因為肇事車輛為小客車的比例極高，因此小客車數與肇事車輛數其實沒有太大的差異，因此**刪除小客車**，而其**他車種大多屬於大型車，因此合併為大型車數量**。

(3) **追撞** (RearEndCollision) :

翻覆變數相較其他事故原因之下，**檢定結果好，因此保留**，而其他事故原因考量到大多數車禍皆為追撞型事故，因此**將起火、自撞和其他事故合併到追撞變數**。

最後刪除溫度、濕度及氣壓計，將原本 37 個變數縮減為 25 個變數。

3.4 模型開發

3.4.1 評估指標

本研究的評估指標分為兩個方面，統計量化指標和誤差值

(1) 統計量化指標

依照實際情況與預測結果的比較，可以分為四種狀態，如表 3，並藉此建構表 4。

實際狀況 / 預測結果	+ (有堵塞)	− (無堵塞)
+ (有堵塞)	TP（真陽）	FP（偽陽）
− (無堵塞)	FN（偽陰）	TN（真陰）

表 3 模型分析結果狀態

評估指標	定義	公式	意義
召回值 (Recall)	真實狀態為陽性且其預測結果亦	$\frac{TP}{FN + TP}$	測試正確識別實際上有堵塞狀況的能力

	為陽性的比率		
精準性 (Precision)	預測為陽性的結果中真實陽性的比率	$\frac{\text{TP}}{\text{FP + TP}}$	測試正確識別會回堵的能力
準確率 (Accuracy)	預測正確占整體的比率	$\frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}$	測試正確識別所有堵塞狀況的能力

表 4 各項統計評估指標

(2) 誤差值

評估指標	定義	公式	意義
均方根誤差 (RMSE)	**預測值與實際值之間誤差的平方和之平均值的根號**	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	**測試預測的回堵里程與實際回堵里程的差距**
低估比例	(模型預測之「回堵里程」＜實際「回堵里程」的資料總筆數佔測試集總筆數的比例	$\frac{n_{(\hat{y}<y)}}{n} \times 100\%$	測試低估車禍嚴重性的發生率

表 5 各項誤差值指標

3.4.2 模型訓練

由盒型圖（圖 6）發現回堵長度變異程度大，不太容易直接預測回堵長度，故本研究想藉由對數函數的單調性質協助我們縮小全距、平滑資料；然而，對於不會回堵的事件無法直接取對數，因此參考文獻後本研究將進行二階段預測。**第一階段目標為預測是否此事故會塞車，第二階段再針對會塞車的事件預測回堵長度**。

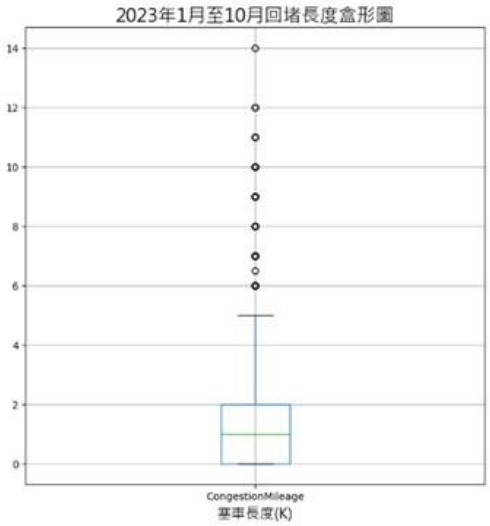


圖 6 回堵里程盒型圖：資料變異程度較大，不容易直接預測回堵數值

下圖 7 為**後續預測回堵里程流程圖**。

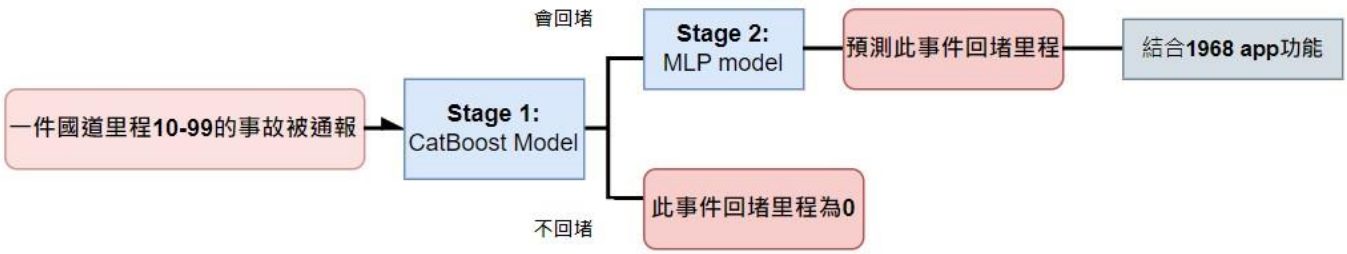


圖 7 預測回堵里程流程圖

第一階段：預測車禍是否會造成回堵（Congestion）

本階段目標是對一未知事件分類是否會造成回堵，訓練集為2023年1月至10月的事件(共5702筆資料)，而測試集為2024年1月至2月(共962筆資料)。模型嘗試SVM、Logistic model、KNN、Random Forest、AdaBoost、XgBoost、CatBoost。初次嘗試的結果如圖七，其中樹狀模型準確率較高，並以CatBoost的表現較佳。

在 2023 年 1 月至 10 月的事件中，若變數 Congestion Mileage 大於 0 則視為會塞車，等於 0 則視為不會塞車，由圓餅圖（圖 8）發現有無塞車的資料較平衡，且針對有回堵的事件經對數函數後，由盒型圖發現降低資料全距及變異程度。

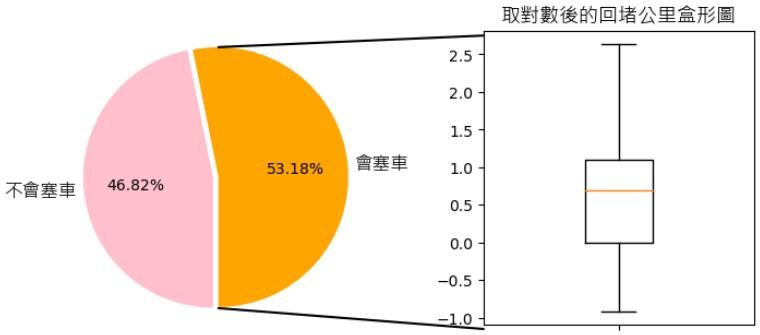


圖 8 區分是否塞車後，對會塞車的數值取對數降低資料變異程度，可比較 圖 6

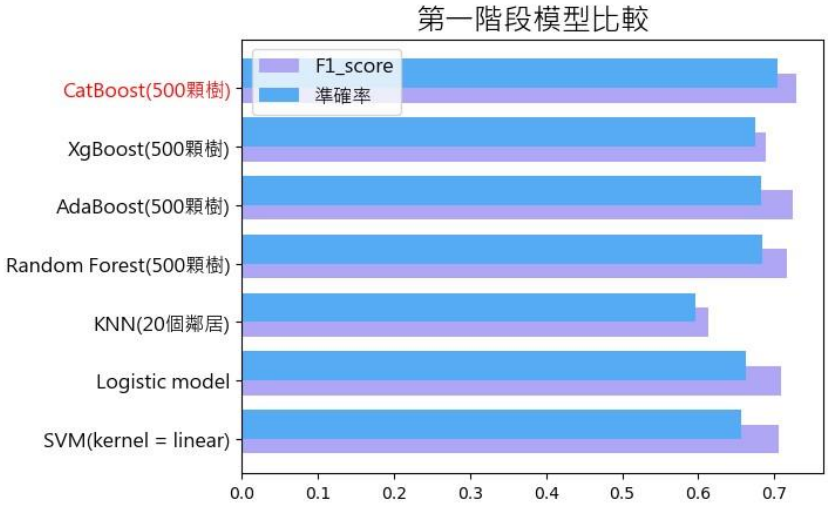


圖 9 第一階段各模型表現：樹狀模型效果較好，其中 CatBoost 模型表現最佳

(1) CatBoost模型介紹

CatBoost 的名稱源於 Category 和 Boost 二字，原 Boosting 的模型對於類別變數，如 TimeCategory 及 DateCategory，需進行 One-hot 編碼的資料轉換，使資料維度變大增加過擬合的風險及計算時間，同時也會忽略類別變數與目標變數的相關，為了解決以上問題，因此有 CatBoost 的引進。

該模型對類別變數有另一套計算方式顯現變數內部的順序性及對目標變數的相關性。不僅能減少維度，也更有效運用類別型資料，且樹狀結構在捕捉交互作用中也相較靈活。

(2) 模型參數選擇

本研究嘗試不同模型參數後，製作出了圖10，由圖八發現準確率幾乎在0.7左右，最高發生在950顆樹且深度為4時，也就是圖中白底的部分，準確率約0.713。

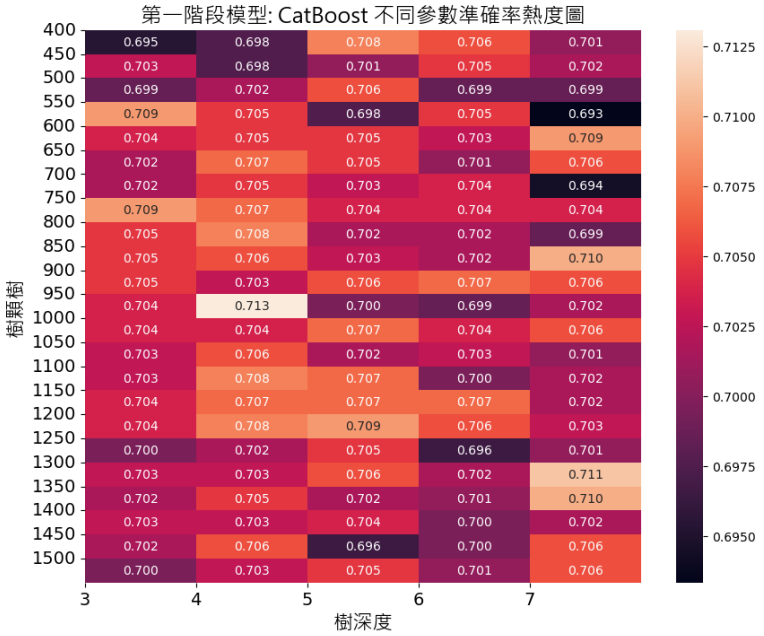


圖 10 CatBoost 模型參數調整熱度圖

(3) 第一階段模型結果討論及調整

模型參數選擇後，使用最高準確率的組合來訓練模型，並製作出 圖11 的混淆矩陣。本研究發現以民眾角度而言，在預測錯誤的 False Positive及False Negative中，False Negative，即**此事件會塞車但卻預測不會塞車較容易引起民眾反彈**，故我們想**調整模型成盡可能網羅確定會塞車的事件**，也就是**提高模型的Recall值為首要目標**。

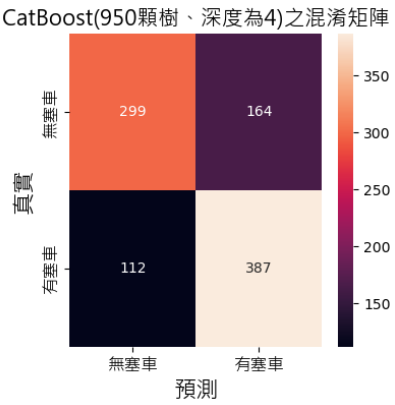


圖 11 第一階段模型混淆矩陣

本研究透過調整在CatBoost的初始訓練樣本的比例達到目的，除了Recall值及準確率作為指標外。F-score亦是可參考的指標（式五），此指標權衡Precision及Recall的關係。當 B=2 時稱為 F2-score，其值的權重會較偏向Recall，故我們選擇F2-score做為另一指標。

$$F_B = \frac{(1+B^2)(Precision*Recall)}{(B^2*Precision)+Recall} \qquad （式五）$$

由 圖13 觀察不同樣本權重下的模型表現，其中選擇F2-score最大處的樣本比

例，即有塞車的事件約0.87而無塞車的事件約0.13，**其Recall值高達0.9以上，由圖12發現False Negative處減少至7筆，代表會回堵的事件有高機率預測準確**。

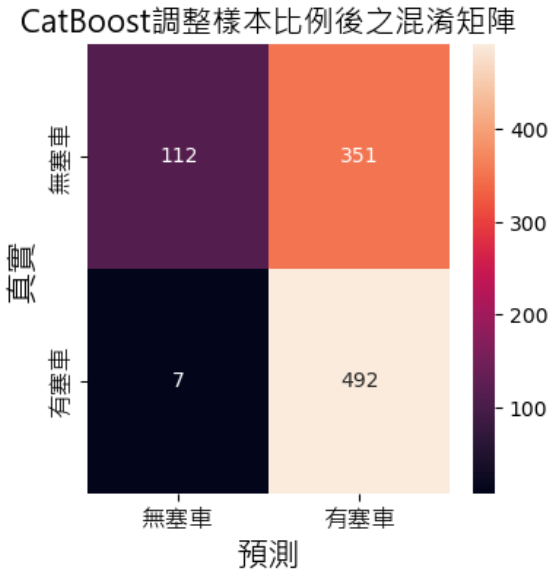


圖 12 第一階段調整後模型混淆矩陣

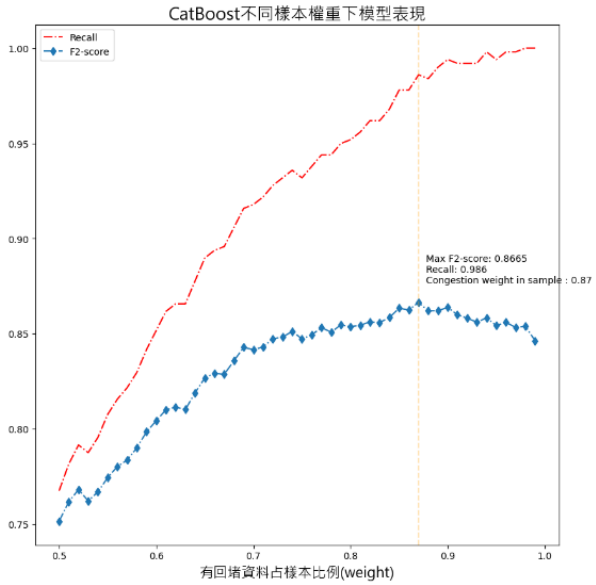


圖 13 模型訓練的兩個指標在不同樣本比例選擇下的摺線圖：選取標準為在維持召回率下，選擇能讓 F2-score 的最大化的樣本比例

(4) 重要特徵討論

Shapley Value可協助計算不同特徵對目標問題的預測有多少貢獻，由圖14得到**以下對於第一階段模型預測較重要的變數**:

1. **事故前平均車速**(Pre_AverageCarSpeed): 其值越低代表回堵長度越長
2. **事故是否波及匝道**(Ramp): 事故若發生在匝道上較不容易造成回堵
3. **時間類別**(TimeCategory): 因為此特徵是類別型變數，圖中可看見有特定時間段對於目標預測有高度貢獻

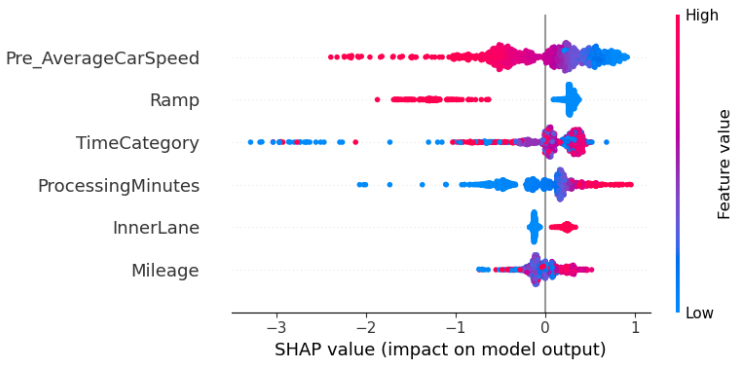


圖 14 變數重要度排序圖：使用 Shapley Value 計算

第二階段：預測回堵里程（Congestion Mileage）

根據前述，此模型的「Congestion Mileage」會取 log，變數選擇則為前述透過各維度縮減方式後所篩選出的 25 個變數。此階段的**訓練集為** 112 年 1～10 月，變數 **Congestion = 1（代表此車禍有造成回堵）的數據**（共 3032 筆資料），而**測試集則為** 113 年 1～2 月，**第一階段 CatBoost 模型預測會回堵之數據**（共 843 筆資料）。

在此階段我們嘗試了線性迴歸模型（Linear Regression）、隨機森林（Random Forest）、支持向量回歸（SVR）、多層感知器（MLP）四種模型進行模型訓練（圖13為各模型比較之結果），本研究模型參數調整的目標為：

極小化 RMSE：希望預測出來的回堵里程與實際回堵里程的差距極小化。

極小化低估比例：預測回堵里程若低於實際的回堵里程，則我們的模型對駕駛（國道主要使用者）來說就沒有效能，故提高效能最好的方式就是降低我們低估回堵里程的比例。

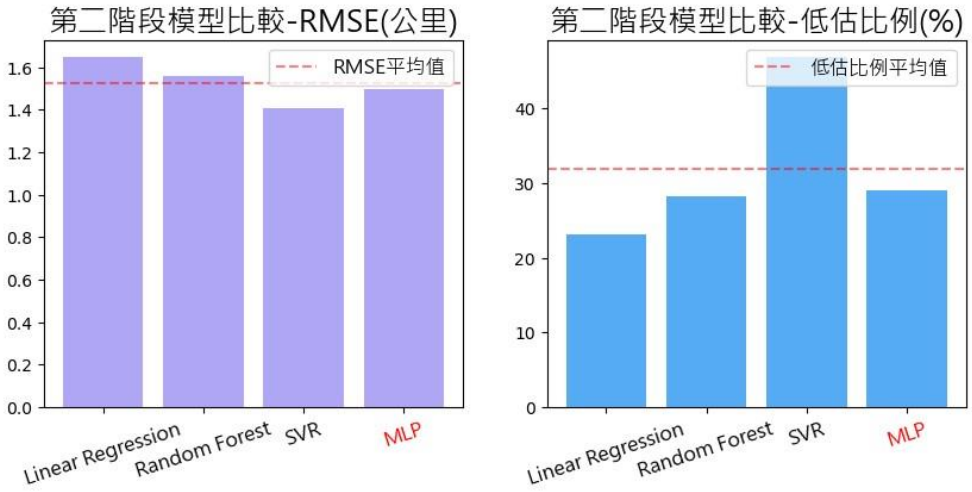


圖 15 第二階段模型比較：MLP 模型相較其他三者在 RMSE, 低估比例上表現都優於平均

由圖 15 可知 **MLP 模型在兩者指標中表現都低於平均線代表其預測誤差、低估比例都較小**，因此最終以 MLP 為第二階段的模型。多層感知機（Multilayer Perceptron, MLP）由多層神經元構成，包括輸入層、隱藏層和輸出層。每層神經元都與下一層的所有神經元相連，這使得 MLP 能夠學習和建模複雜的非線性關係。此階段建構了包含輸入層，六層隱藏層（分別為 256、128、128、64、64、32 個神經元）及輸出層的 MLP 模型，並使用修正線性單元（ReLU）激活函數來引

入非線性，使得神經網絡能夠學習複雜的數據。在隱藏層之間加入 Dropout 層，隨機棄用部分神經元，減少過擬合的風險。最後使用 Adam 優化器來加速學習過程與調整學習率。在訓練過程中，模型進行了 150 個訓練週期，並使用批量大小為 128 的方式來更新模型權重。訓練完成後，將對數轉換後的預測結果還原至原始尺度，以便與真實值進行比較。

根據結果，初步猜測 MLP 模型表現優於其他三種模型的主要原因為：

1. 非線性建模能力：MLP 能夠捕捉輸入特徵與目標變量之間的複雜非線性關係。在面對特徵包含交通因子，施工因子，天氣因子等等多變且非線性的數據時，能夠進行更精確的預測。
2. 多層特徵學習：MLP 透過多層隱藏層來學習數據中的深層特徵表示，這種特徵學習的能力遠超過單一層次的模型（如線性迴歸）。
3. 過擬合的處理：在模型中適當加入Dropout層能夠有效減少過擬合的風險。
4. Adam 優化器：通過自適應的學習率調整策略，能夠更有效地找到最佳權重，對於模型的快速收斂和穩定性有極大幫助。

對於第二階段的 MLP 模型，**測試集為第一階段被預測會回堵的數據**時，約 29.024% 的預測回堵里程比實際值低，代表模型低估了回堵的嚴重程度。而 **RMSE** 表示模型每次預測的回堵里程和實際里程之間的誤差平方平均開根號**約為 1.496 K**。

4 研究結果與討論

4.1 研究結果

4.1.1 二階段模型效果

將訓練完畢的兩模型結合後，測試集的 RMSE 為 1.45K，被低估比率為 24.94%。我們認為誤差約 1.45 K 在國道上是可接受範圍。然而，模型對低估的訓練仍有進步空間，因第一階段 Recall 值已夠高，故主要原因在第二階段模型，未來應對該階段做更細緻的建模以降低低估比例，目前仍以減少誤差為準則。

4.1.2 重要變數

由圖 14 發現**重要的因子主要為**，

(1) **事件因子**: 事故發生位置、時間點、處理時間、里程數。

(2) **事故前路況**: 事故前平均車速。

4.2 誤差原因探討

1. **特徵缺乏精準度**：

在國道上因車速快，缺少一分鐘會損失不少有效數據，進而導致模型效果不佳，例如 ：

(1) 由圖14的特徵重要度資料中最高為「事故前平均車速」，此特徵需整合單一事件前四小時的門道所有數據，並精準地提取事故前十分鐘的數據以計算車速，每一事件至少耗時3分鐘計算。然而，另一特徵「事故前十分鐘的車流量」因原始數據是以每五分鐘為時間段，使得數據不夠精準對模型預測效益大減。

(2) 天氣資料本研究參考文獻後認為事故前降雨量會是重要特徵，然而降雨量的歷史即時資訊難以尋找。此外也無對應國道的天氣蒐集站，本研究只能利用國外天氣網站，並從地圖尋找較近的觀測站資料使用，其中甚至沒有降雨量資料。因此**本研究認為需針對預測事故蒐集相關數據，例如: 每分鐘車流量、每分鐘降雨量等……**

2. **國道門道故障**:

在特徵擷取過程中發現門道有故障的風險，**建議能對門道故障設計備案**以避免長時間段的資料遺漏。

3. **事故簡訊內容**:

(1) 在拆解簡訊內容時，我們發現許多定義不明確或是車種名稱紊亂的情形，例如: 工工程車和施工車、水泥車和預拌混泥土車、小自客_x000D_\n和小自客\n等…… 使得我們只能自行分類。

(2) 本次研究目標為預測回堵長度，然而有至少100個事件的回堵狀況遺失。

本研究建議能對簡訊訊息設計固定的格式，且對車種的分類明確，才能降低特徵工程的難易度並增加模型的精準度。

5 實際應用與可行性

基於目前高公局的各項道路監控及通知系統，道路監控包含天候、事故、車輛偵測等，而通知系統則包含高速公路 1968 App 和網站、路旁的資訊可變標誌、即時路況電台和警察廣播電臺等，本次研究所使用的資料大部分來自高公局提供，只有少數天候資料因為氣象局無法查詢事故時的及時天氣而使用其他網站，但高公局的道路天氣監控即可補足這些資料，因此後續**無論是想對其他路段進行模型建構和預測，或是定時更新預測模型，甚至擴充資料和變數加強模型效果都不會太繁瑣**，只需要調用故有的資料即可執行，因此可行性相當高。

實務應用方面，考量到目前有效的通知系統和研究成果，**1968 App 為複合式應用程式**，內建的許多功能皆**可以和本研究結合**，做出更好的路況評估，以下為 1968 App 現有功能與可改善方式建議：

1. **北中南等各區的及時路況及路網圖**（參考圖 16）：

目前此功能主要是針對當下各路段的即時車速，更新率為一分鐘，如果能在事故發生的當下對該路段接下來可能的壅塞情況做預測，並在較嚴重的區域**提早警示可能回堵公里數，即可提供第一時間的視覺化通知**，讓副駕駛或尚未上路的駕駛自行評估使否更改路線，減少後續更多用路人湧入該壅塞路段。

2. **路線時間預測、訂閱路段推播**（參考圖 17）：

此二功能與 Google Map 的定位計算旅途時間一樣，可以依照當前路況或訂閱的時間路段路況，評估選定的起始與目的地交流道需要多少時間抵達，事故發生時，**透過本研究的模型在第一時間更新各位置即將造成的回堵程度，即可增加時間預測精準度，並減少預測偏差**，防止用路人誤判行車時間。

3. **路況事件推播**（參考圖 18）：

此功能開啟後，定位會依照目前車輛位置，對前方特定公里內發生的事件進行語音推播，且駕駛可自行選擇想接收的通知，如管制事件、施工、事故等，在事故類別的**推播內容中可以加入預測回堵公里數，以及回堵範圍前的交流道口，提供駕駛迴避壅塞路段的交流道選項**，有助於減少壅塞路段的部分車流，避免壅塞加重。

圖 16 即時路況與路網圖：可在事故第一時間標註地點、排解情況、預測回堵等資訊供用路人參考

圖 17 路段時間預測：在預測交通時間上可增加當前是否有事故影響，並提供行車時間差異進行比較

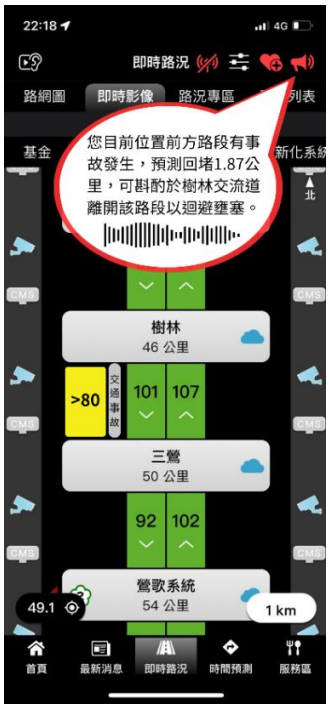


圖 18 路況事件推播：可以根據用路人位置，語音通報前方事故的預測回堵資訊

6 參考文獻

1. 國道易壅塞路段(https://data.gov.tw/dataset/33191)
2. Abdi, A., Seyedabrishami, S., & O’ Hern, S. (2023). A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data.Journal of advanced transportation,2023(1), 8737185.
3. Tseng, F. H., Hsueh, J. H., Tseng, C. W., Yang, Y. T., Chao, H. C., & Chou, L. D. (2018). Congestion prediction with big data for real-time highway traffic.IEEE Access,6, 57311-57323.
4. Timeanddate(https://www.timeanddate.com/)
5. 高工局交通資料庫(https://tisvcloud.freeway.gov.tw/history-list.php)
6. 113 年國道智慧交通管理創意競賽資料下載(https://freeway2024.tw/links#links)

交通部高速公路局

113 年國道智慧交通管理創意競賽

投稿作品 3 年內是否公開獲獎切結書

立切結書人_____等參加高速公路局舉辦之113年國道智慧交通管理創意競賽所投稿之作品

☑未曾於 3 年內公開獲獎

☐曾於3年內以_____（作品名），公開參加_____（競賽），獲得_____（獎項），現投稿之作品，與之前獲獎作品有顯著之差異性，且已於報告中敘明其前後差異，若經主辦單位審核發現參賽作品有違反本比賽規則所規範事項者，本人同意被取消參賽資格；如已得獎，亦同意被追回已頒發之獎項，絕無異議。

立切結書人：(須全體成員簽章)

姓名及簽章	身分證字號	聯絡電話及戶籍地址
陳昭鈞	F131566619	0902122828，新北市三重區力行路二段178號5F
郭依璇	A230575317	0979438843，臺北市文山區景福街76号11樓
李翊業	F125869002	0928041090，屏東縣屏東市廣東路838巷18號
林琦珈	F309556501	0958200612，臺北市大安區龍安里1鄰和平東路一段199巷5弄2-1號

中 華 民 國 113 年 6 月 23 日