作品名稱：針對高事故發生地區之事故預測模型

隊伍名稱：Civit

摘要

隨著國道的使用人數上升，事故發生的次數也漸漸上升，而事故一旦發生，常常造成財物損失以及重大傷亡。根據統計結果，108 年共發生 31,218 件國道事故，相較於 107 年 (24,809 件) 有明顯上升，A1 類 72 件 (107 年 63 件)、A2 類 2,076 件 (107 年 1,770 件)、A3 類 29,070 件 (107 年 22,976 件)，可以發現不管 A1、A2、A3 事件都有上升的傾向，而三大肇事原因則和去年的順序相同，依序為「未保持行車安全距離 (42.53%)」、「未注意車前狀態 (19.26%)」及「變換車道或方向不當 (12.08%)」。

透過分析各個國道可以發現國道一號所發生的事故數量要比其他國道發生事故數量多，且特定地點發生事故的數量較其他地點高，因此優先分析國道一號”01F0339N/S” 測站周遭為何會導致事故頻繁發生，並針對此地點建立預測模型。以交通部高公局提供之”108 年國道事故資料” 做探索式資料分析，根據其結果於 108 年 TDCS 資料庫中收集變數，透過收集到之變數來預測該時段是否發生事故，且找到對此預測較有貢獻的變數，藉由這些變數試圖去解釋一些現象。最終我們可以將此模型推廣至每一個國道測站，並將此模型與 APP(例如高速公路 1968) 聯動，在預測發生事故時即時提醒用路人。

本分析使用的資料有”M05A-SpaceMeanSpeed”(站間各車種平均行駛車速)、”M05A-交通量”(站間各車種於五分鐘內經過之車流數量) ”M07A-交通量”(各車種以某測站為起點之車流數量)、”108 年國道事故資料”、”108 年北區散落物資料”。

本文的內容主要分為三部份，第一部份會做一些基本的探索式資料分析，並點出四個常發生事故的地點，並聚焦於事故發生頻率最高的測站”01F0339N/S”，最後推敲對該處有用之變數。第二部份會根據前一個部分所得出的結論進行統計建模以及模型預測。第三部份會做模型的總結以及討論。

Contents

1 探索式資料分析

本篇章針對不同變數進行探索式資料分析，並分別根據時間、空間、肇事進行探討，同時在分析過程中選擇預測變數以利後續建立模型。

1.1 資料前處理

1.1.1 遺失值

先將 108 年國道事故資料資料載入，發現在此資料中存在一些遺失值，於是針對遺失值進行合理的填補。

- 公尺：遺失 19 筆，以 0 補
- 肇因：遺失 13 筆，以"43：不明原因肇事" 補
- 車種：遺失 1532 筆，以"G06：其他車" 補

1.1.2 變數合併及分解

- 將" 發生-路線-公里" 以及" 發生-路線-公尺" 合併成資料型態為小數的" 公里"
- 將" 發生時間" 分成" 發生時間-時"，" 發生時間-分" 這兩個變數

1.1.3 M05A-SpaceMeanSpeed 資料錯誤問題

將 M05A 資料載入，取變數"SpaceMeanSpeed" 以及" 交通量" 做後續備用，對"SpaceMeanSpeed" 變數篩選高於 200 的觀察值，可以發現 "SpaceMeanSpeed" 資料有部分出現問題 (主要集中於 6/25 日)，其中速度至 200 以上不合理[1]，所以我們將這些觀察值都設為 NA，同時，假若一樣本交通量為零，速度就會被系統設定為零，但事實上沒有車經過就沒有速度，所以也將這些速度的數值設為 NA[2]。接著進入探索式資料分析。

[1]如 2019/6/25 日 1:00 就有一個速度為 6660 的資料，其中還有更多高於 1000 之數值

[2]若未有車輛經過，且速度被設為 0，對其取平均將會使得估計不準確

1.2 依變數進行事故統計

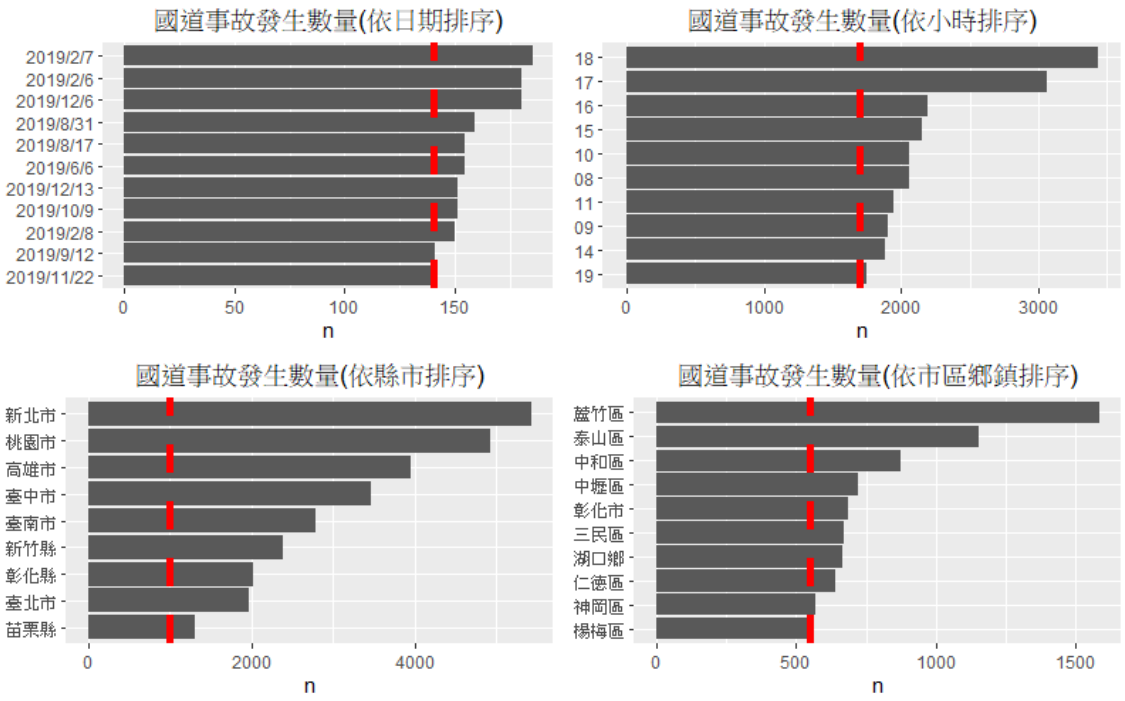


Figure 1: 事故發生數量比較 (依變數比較)

將事故發生數量依各種不同類別變數分類。其中紅色虛線為該類別變數排名第十之事故發生數量。

1. 依日期計算事故發生數量，可以看到雖然事故發生數量前三高的日期皆為年假，但沒有和非假日 (8/17，11/22) 差異太多。

2. 依小時計算事故發生數量，可以看到事故發生於上下班時段的數量遠高於非上下班時段數量，我們認為是否為上下班時段和發生事故應該有關係，故可考慮作為變數加入模型。

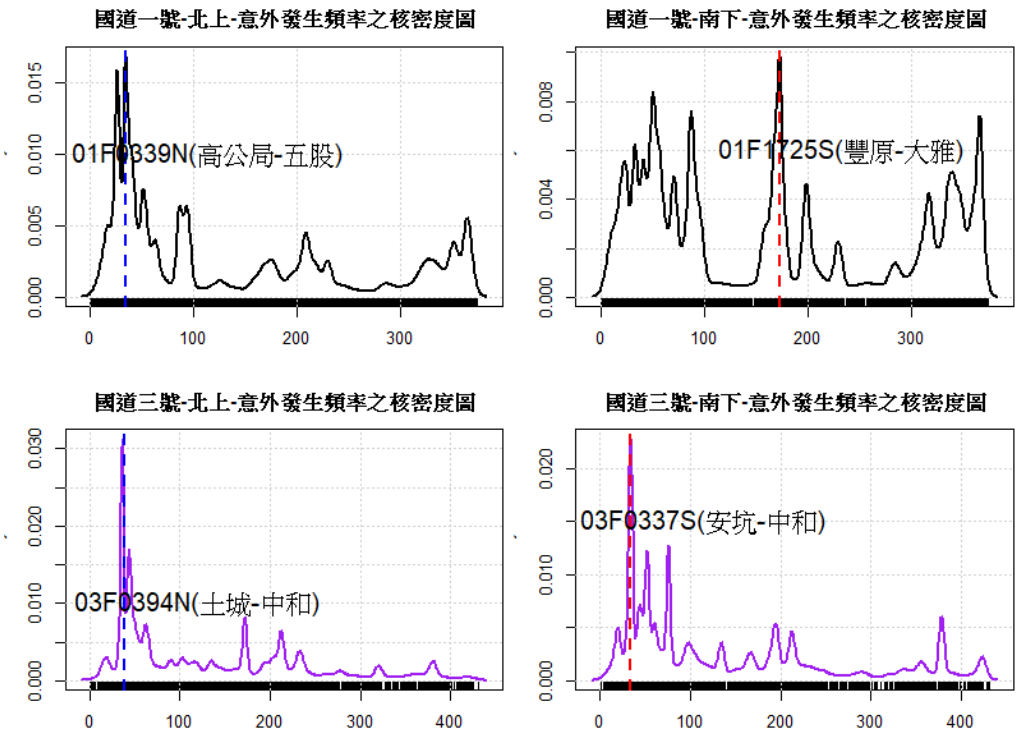
3. 依縣市計算事故發生數量，可以看到新北市以及桃園市的事故數量最多，接著是高雄市以及台中市。

4. 進一步統計主要發生在哪些鄉鎮，可以看到蘆竹區 (桃園市) 以及泰山區 (新北市) 遠遠高於其他鄉鎮，其中在這兩個鄉鎮發生之事故主要的來自國道一號 30K 至 55K。

由此可知，不同地點、不同時間之事故發生頻率不同，且肇因也不盡相同。接下來會依據”地點”、”時間”、”肇因”三個面向進行探討。

1.3 地點與事故間之關係

針對不同國道統計 2019 年的肇事數量，可以發現國道一號所占比例最多 (20986)，接著是國三 (7645)，接下來分別是國道二號以及國十，但數量皆未超過 1000，將重點放在國道一號以及國三上，將國道一號以及國道三號分為南北向，透過變數：” 公里” 繪製核密圖 (density plot)，圖中縱軸數值反應對應公里處事故發生密度高低。



我們由左上角依序介紹至右下角。

1. 國道一號北上 0 至 100 公里處的意外事故發生頻率較高，在”01F0339N” 測站左右時意外事故發生頻率達到最高。

2. 國道一號南下 0-100 公里處，170-200 公里處，以及 300 公里以後之事故發生的頻率較高，在”01F1725S” 測站左右時意外事故發生頻率達到最高。

3. 國道三號北上 0 至 100 公里處的意外事故發生頻率較高，且在”03F0394N” 測站左右時意外事故發生頻率達到最高。

4. 國道三號南下 0 至 100 公里處的意外事故發生頻率較高，且在”03F0337S” 測站左右時意外事故發生頻率達到最高。

整體來說，國道一號相較於國道三號有更多的事故發生量，且國道一號又屬測站”01F0339N/S” 附近所發生之事故最多。為了不讓重點失焦，我們將專注於國道一號 30K 至 55K(含南北向) 處進行探討[3]

1.3.1 國道設施以及事故發生次數之關係

觀察國道一號 30K 至 55K 發現事故發生和國道路線設施 (交流道、服務區) 有高度相關，以國道一號以及國道二號為例子，對” 公里” 變數繪製核密圖，並將將有設施的公里數標出 (紅色三角形)，其結果如下圖

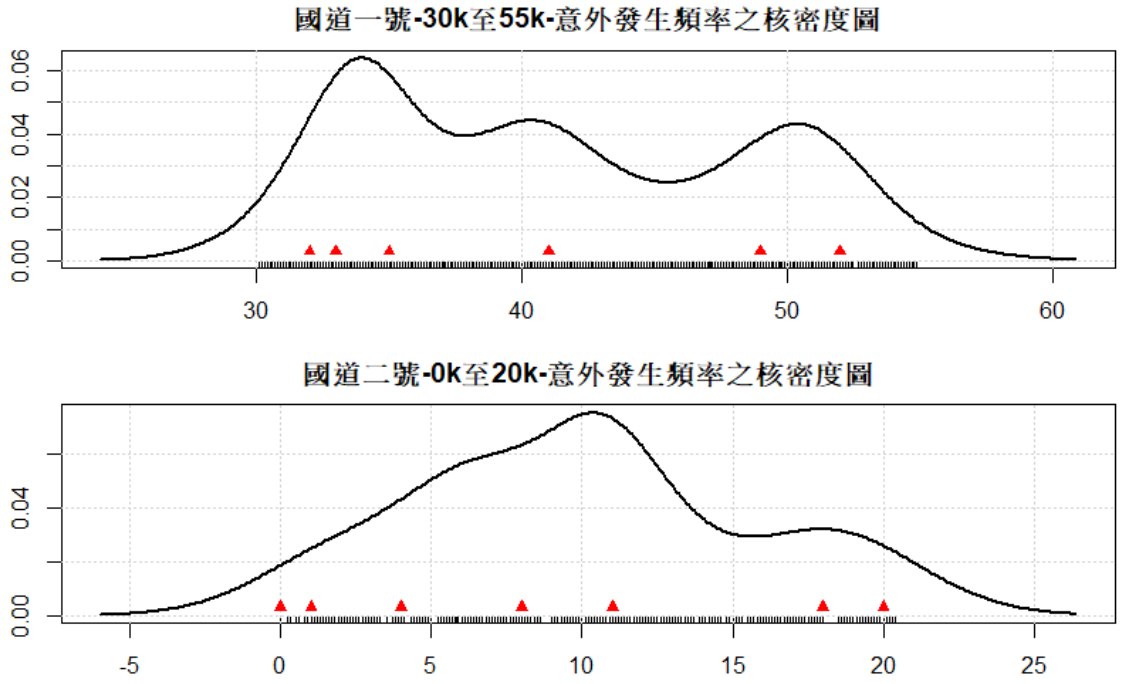


Figure 2: 事故發生數量比較 (加入國道設施)

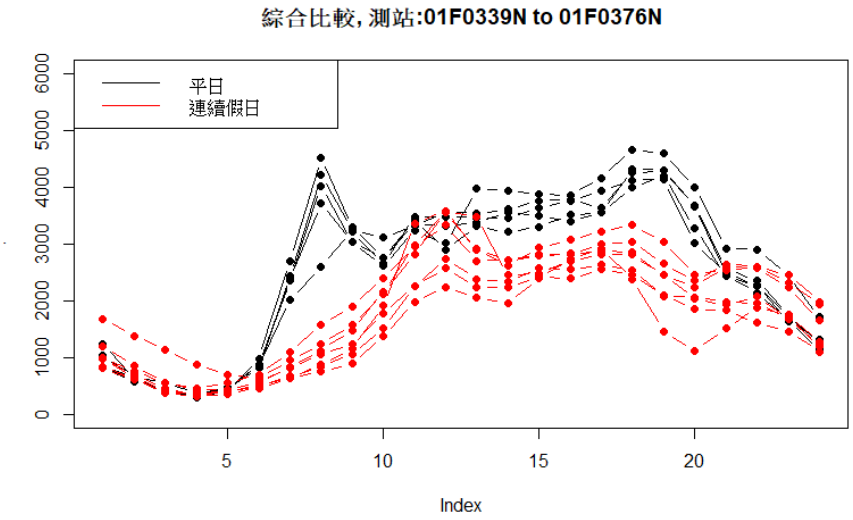
可以發現，若路段設施越多 (路況越複雜)，事故發生數量越高。因此認為若有由該設施出發的交通量數據 (M07A 之交通量)，或許能對事故預測有幫助。

[3]事後我們也有針對”01F1725S”、”03F0394N”、”03F0337S” 測站進行預測，其各方面數值都有不錯之結果, 間接說明此模型的泛用性。

1.4 時間與事故間之關係

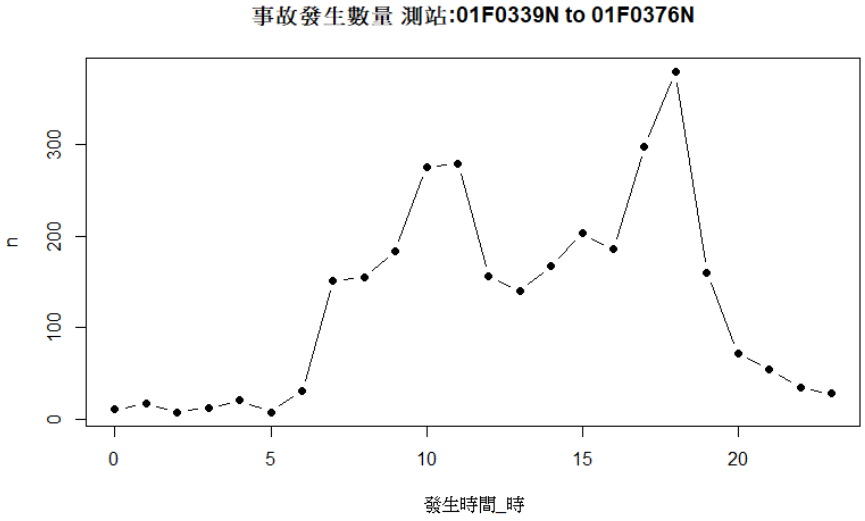
1.4.1 國道車流變化與事件發生之相關性

針對測站"01F0339N" 至"01F0376N" 收集 6/10 至 6/14(非假期) 的車流資料，依照事故發生時段計算車流量並作圖，發現平日與平日之間的車流相似，如下圖四黑線部份。同時針對同一個測站收集連續假日 (以 2/2 至 2/8 為例) 之車流量，並依相同方式做圖，得到的車流趨勢如下圖紅線部份，發現連續假日與連續假日之間的趨勢也十分一致，然而連續假日和非連續假日之間的趨勢有些許差異，連續假日的車潮較非假日的車潮晚到來 (約在 11 至 12 點車流達到高峰)。



我們發現連續假期與非假期車流趨勢有差異，故連續假期與非假期應該分別討論，並納入預測變數中。

而車流量是否和事故數量有相關，答案是肯定的，我們接著看到下圖四，可以看到隨著車流變化發生事故的數量也隨之變化。



1.5 肇因與事故間之關係

接著我們將站在三個立場分析事故類型，分別是社會整體角度、肇因角度以及車種角度。

1.5.1 社會整體角度

站在社會整體立場，能夠順利完成一趟旅行，平安通勤不遲到是大家的共識，而這些共識都取決於是否發生事故。為達成此共識就必須列出常見肇因並加以宣導。因此對肇因進行統計並列出前三高的肇因為。

1. 未保持行車安全距離
2. 未注意車前狀態
3. 變換車道或方向不當

站在社會整體利益的角度，應該要加強宣導：保持行車安全距離、注意車前狀態、變換車道前注意路況等等。若政策執行方向根據此順序決定對於社會整體的效用較高。

1.5.2 用路人角度 (肇因)

若站在用路人自身的角度，或許十次未保持安全距離，還未發生事故。如果用路人抱有這種僥倖的心態，他接下來也會僥倖認為這次不會

發生問題；若用路人因為某原因發生事故了，他受傷的機率為何？條件在因某肇因發生事故下，計算用路人會受傷或死亡的比率為何，藉此透過比率來估計機率[4]後，對比率進行排序，可以得到下方表格。

肇因	事故發生總數	受傷或死亡事故總數	比率
違反特定標誌 (線) 禁制	41	17	0.414634146
拋錨未採安全措施	32	13	0.406250000
酒醉 (後) 駕駛失控	110	38	0.345454545
疲勞 (患病) 駕駛失控	79	23	0.291139241
未依規定減速	101	25	0.247524752
超速失控	25	6	0.240000000

可以發現，比率最高的為”違反特定標誌 (線) 禁制”，若該用路人違反特定標誌 (線) 禁制且發生事故，那受傷或是死亡的比率有 0.41，大約為四成，其他的解釋方式以此類推。透過回報此種比率可以向用路人提醒，若抱有僥倖心態違規，且運氣不好發生事故，其將會受到巨大損失 (受傷或是死亡)，若政策執行方向根據此表決定對於個人的效用較高。
最後透過該表可以得出以下結論。

- 加強拋錨之安全措施宣導
- 嚴禁酒駕以及疲勞駕駛

1.5.3 用路人角度 (車種)

計算條件在事故發生的狀態下，哪一種車發生事故後會造成人員受傷的比率最高，其結果呈現於下表格。同時將事故數量小於 20 的車種移除。

[4]在這裡我們將事故發生數量低於 20 的肇因排除，第一個問題是：樣本不夠大可能會造成真實比率估計不準確，第二個問題是：它可能是很罕見的肇因，譬如 2019 年吸食違禁物後駕駛且發生事故之數量為 1，且其為 A1、A2 類交通事故，假若我沒有將此移除，會被誤解為吸食違禁品發生事故有很高機率死亡，但這是一個較稀有的事件 (rare event)，所以發生次數不多，導致估計不準。除此之外若倡導大家不吸食違禁品上路效益並不高，因為用路人吸食違禁品上路的比例很少。

車種	事故發生總數	受傷或死亡事故總數	比率
機車	23	19	0.82608696
民營客運	123	22	0.17886179
自用半聯結車	82	10	0.12195122
營業用半聯結車	33	4	0.12121212
自用曳引車	43	5	0.11627907
計程車	340	38	0.11176471

撇除誤上國道的機車，可發現比率較高的車種主要為大型車，而計程車緊接在後。若考量到個人效用，可以根據此表格之結果對特定車種進行宣導來降低的危險性。且可以發現，不同車種於特定時間點的交通量之變化似乎會影響到事故發生機率，故在 M05A，M07A 中同時考慮各個車種的交通量並放進變數欄中。最後透過該表可以得出以下結論。

- 機車誤上國道十分危險
- 大型車以及計程車請小心駕駛

2 模型預測

透過上面的整理，收集了 M05A 以及 M07A 中的交通量 (依車種分別)，該日是否為節日以及該時段是否為上班時段的變數，除此之外，額外加入" 是否有掉落物存在" 以及 M05A 中的"SpaceMeanSpeed"。將資料下載完成後，我們希望透過上述資料整理出 2019 年每天每小時下的各種變數。舉例來說收集樣本："2019 年 1 月 1 日 12 時" 下交通量為多少，平均車速是多少，是否有掉落物，是否於此時間區間發生事故等等，接著再將此做為預測模型並預測。

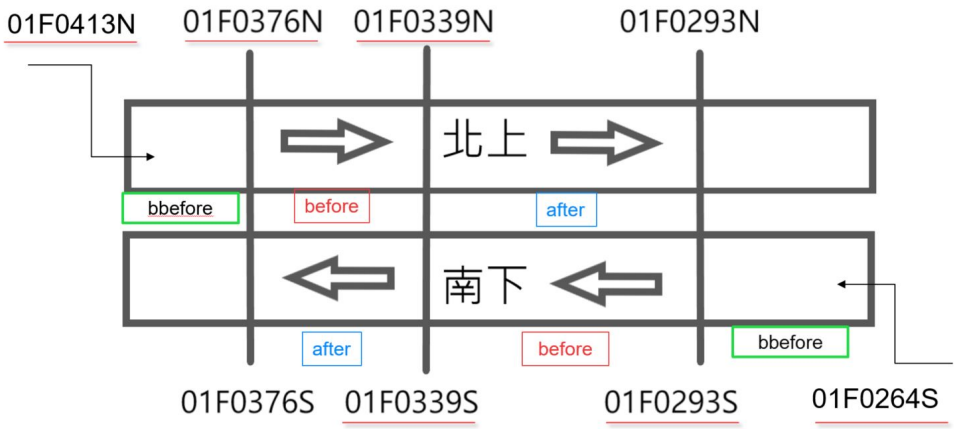
2.1 變數整理過程

- **step1**：將掉落物的資訊載入，並將遺失值全部移除，選擇掉落物介於國道一號北上南下 29K 至 38K 之間的數值，記錄下處理結束之時間 (準確至小時) 並根據其為北向或是南向分類，若有掉落物則設定為"yes"，反之為"no"。
- **step2**：將 M05A 的資料載入，將起始測站為"01F0339S"、"01F0293S"、"01F0264S"、"01F0339N"、"01F0376N"、"01F0413N" 的資料收集下來，分為南北兩向。而 M05A 的資料每五分鐘收集一次，為了將 M05A 的資料與 M07A 配對，將 M05A 資料中同一小時之"SpaceMeanSpeed" 取平均，" 交通量" 做加總，如此一來 M05A 就和 M07A 同為以小時為單位的資料。
- **step3**：將 M07A 的資料載入，將起始測站為"01F0339S"、"01F0293S"、"01F0339N"、"01F0376N" 的資料收集下來分為南北兩向，將 M05A 以及 M07A 依日期時間合併。
- **step4**：將合併後的資料對" 車種" 變數做 pivoting。針對某一欄類別變數進行延展，將整個資料矩陣變得更加寬廣。示意圖如下。

date + hour	type	speed
2019/01/01 00	31	1
2019/01/01 00	32	2
2019/01/01 00	41	3
2019/01/01 00	42	4
2019/01/01 00	5	5
2019/01/01 01	31	6
2019/01/01 01	32	7
2019/01/01 01	41	8
2019/01/01 01	42	9
2019/01/01 01	5	10

date + hour	speed 31	speed 32	speed 41	speed 42	speed 5
2019/01/01 00	1	2	3	4	5
2019/01/01 01	6	7	8	9	10

- **step5**：可以注意到，若是北上車流，則必定會先通過”01F0413N”，接著是”01F0376N” 以及”01F0339N”；同理若是南下車流，則必定會先通過”01F0264S”，接著是”01F0293S” 以及”01F0339S”。所以將起始測站名稱”01F0264S”、”01F0413N” 改為 bbefore，將起始測站名稱”01F0293S”、”01F0376N” 改為 before，起始測站名稱”01F0339S”、”01F0339N” 改為 after。接著再對” 起始測站” 做 pivoting。



- **step6**：將小時為早上 7、8、9 點以及下午 17、18、19 點的資料標註為上班時段，其餘則否。
- **step7**：將日期為適逢年假、228 連假、清明節、端午節、中秋節、國慶日的資料標註為連續假期，其餘則否。
- **step8**：將 108 年國道事故資料中事故位於國道一號 29K 至 38K 之間所發生之事故收集下來，並記錄時間。
- **step9**：將上述所有資料根據日期、小時連接在一起，最後將日期以及小時移除。

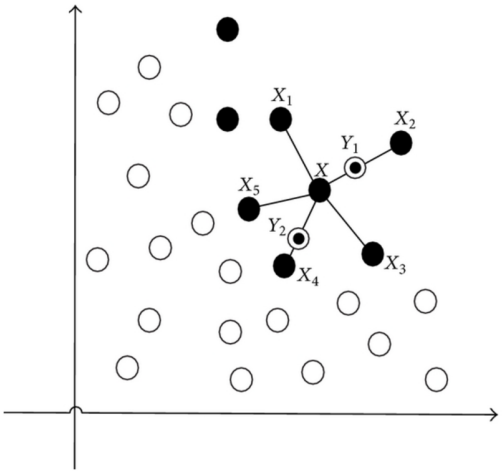
最後資料整理完將會是一個 17520*45 維度的結構 (365 天 * 24 小時 * 2 方向 = 17520，45 個變數)。

2.2 套用 XGBoost 模型

將資料整理好後，將資料依 7：3 的比例切分成訓練集 (training set) 以及測試集 (testing set)，其中訓練集顧名思義就是要讓我們訓練模型的資料集，而將資料訓練完成後我們將會在測試集上面看看訓練成效。將訓練集資料套用至 XGBoost 模型[5]，測試集的資料將放到整個模型配適結束後才會使用到。

2.2.1 依變數比例不平衡

若成功將資料收集，可以發現訓練集中大部分的時段都未發生事故，只有少部分時段有發生事故，若不對此作預處理，模型將無法分辨發生事故與否，會傾向全部預測未發生事故，因為這樣能帶來的準確率就很高了。這種問題被稱作 imbalance data 的問題，其中解決方法有 down sampling、over sampleing、ROSE 以及 SMOTE。這裡我們選用 SMOTE 來解決此問題，因為其不需要分佈假設也不會造成資訊喪失。SMOTE 的概念很簡單，將資料視為存在一多維度空間中的樣本點，在較少的那個類別中，任選兩點並將其連接成一條線，其中隨機選擇線上一點作為新的觀察值。透過此方法不斷生成資料直到資料平衡，其示意圖如下。其中 Y1，Y2 就是透過此方法生成之新樣本點。



[5]配適三個模型：logistic+boosting、random forest、XGBoost。屬 XGBoost 效果最好，故於本文中節錄 XGBoost 的內容

2.2.2 參數選擇

將資料套用至 XGBoost 模型，此模型可以調控的參數很多，而針對每筆不同的資料，參數大致上都不太一樣。現在我們需要透過一些準則去選取參數，其中準則有很多選擇，大致上有準確率、Sensitivity、Specificity、PPV、NPV 這五個統計量，而這些準則的算法如下圖所示。

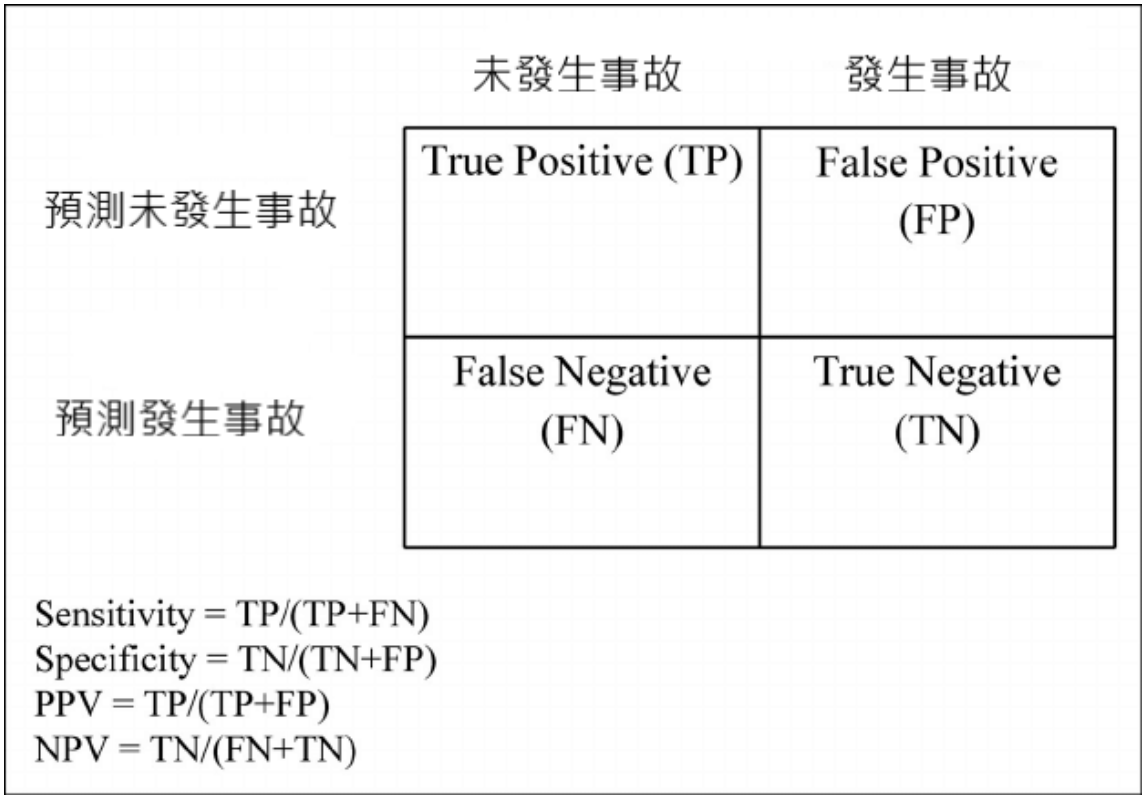


Figure 3: 統計量之計算方式以及意義

我們希望真實發生事故下，模型真的有預測到發生事故的比例越高越好，也就是我們需要注重的統計量為 Specificity。除此之外也希望預測未發生事故下真實未發生事故的比例越高越好，也就是若有較高的 PPV 那將會對模型加分。總而言之，我們希望預測未發生事故下發生事故的數量越少越好。

選好準則 (Specificity) 後我們透過十折交叉驗證 (10-fold cross validation) 的方式選擇會帶來最高 Specificity 的參數。將訓練集的資料切分成 10 份，分別命名為 fold1 到 fold10，取 fold2 至 fold10 作為訓練集中的訓練集，同時 fold1 作為訓練集中的驗證集，在 fold2 至 fold10 裡嘗試不同變數組合，並在 fold1 上衡量哪個變數組合帶來的 Specificity 會最高。接著變更訓練集中的驗證集為 fold2，訓練集中的訓練集變更為 fold1 以及 fold3 至 fold10 後重複先前的動作 10 次，對所有 fold 的 Specificity 取平均後選擇對應到最高 Specificity 之變數組合作為最後變數，示意圖如下。

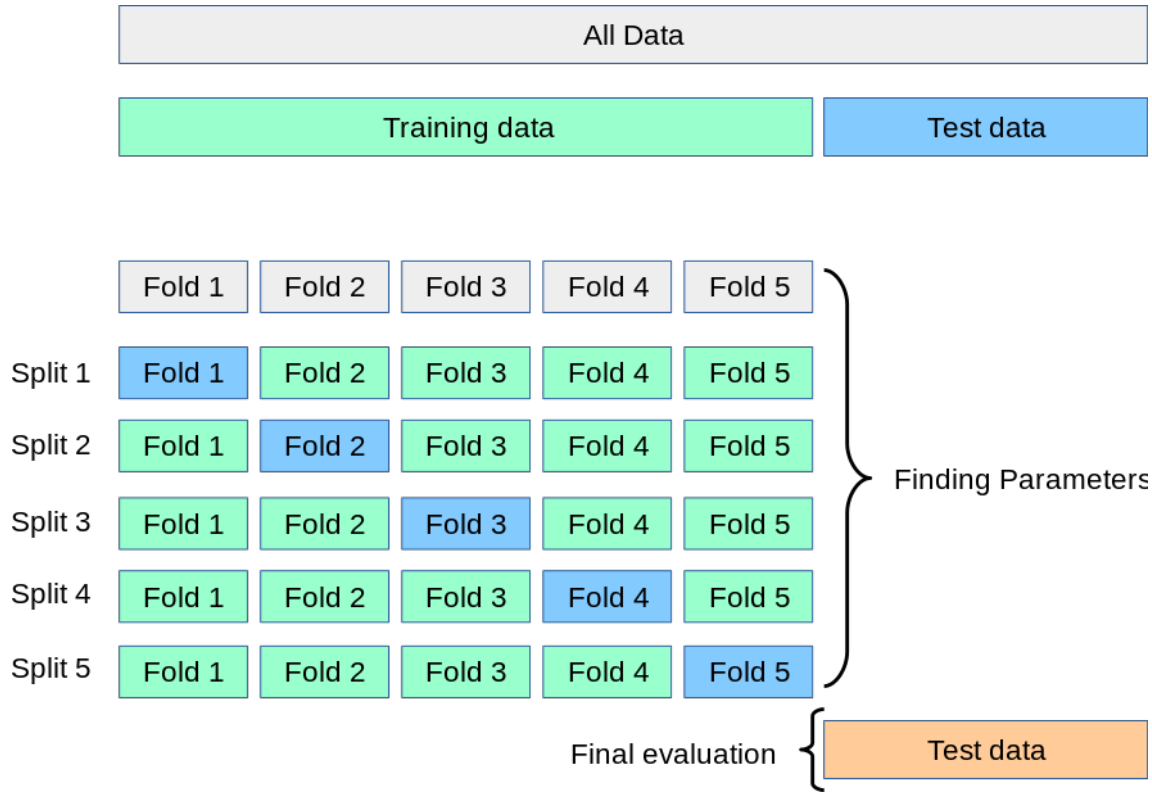


Figure 4: 五折交叉驗證示意圖

最後每個 fold 中我們嘗試 30 個不同的參數組合，完成模型配適。

2.2.3 配適效果

將模型配適完成，計算該模型在測試集上的表現，其結果如下圖。

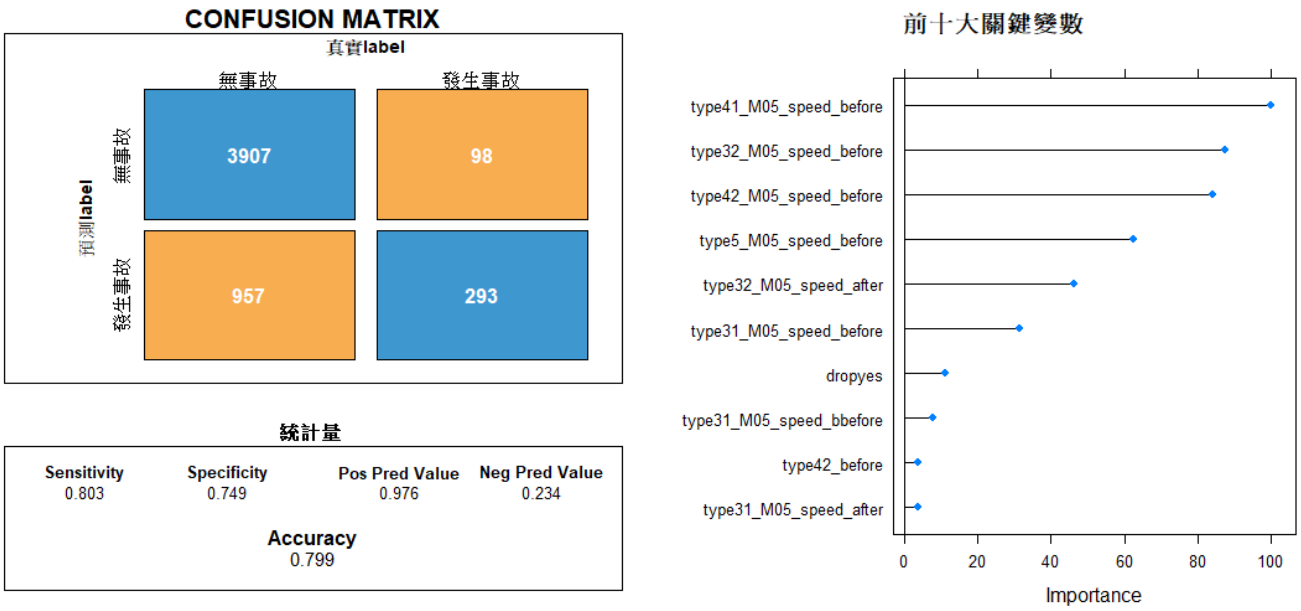


Figure 5: 混淆矩陣

可以看到 Specificity 達到 0.749，且準確率達到 0.799，PPV 達到 0.976，表現算是不錯。接下來看到重要變數的部分，重要變數的獲得方式是透過計算變數被使用的次數[6]。就結果而言，可以看到 M05A 資料中，各種車種的速度是一個很重要的變數，且是否有掉落物也是一個重要指標[7]。近年來，有更多方法探討 XGBoost 模型之解釋性，例如 SHAP value，我們將於後面章節一並進行探討。

[6]XGBoost 是由許多分類樹模型，透過 Boosting 的方式整合在一起，每次 Boosting 就會建立一個分類樹模型，透過計算 A 變數於 n 次 Boosting 中被使用了幾次就可以得到該變數之相對重要性，當然其背後道理又更加複雜，涉及一些為防止過度擬合而產生之超參數 (Hyperparameter) 估計，這裡我們就不深入探討

[7]將模型配適於"01F1725S" 以及"03F0394S" 測站可以發現主要的重要變數都和各種車速有關，而位於國道三號的"03F0337S" 測站有不同的重要變數。"03F0337S" 測站的重要變數主要為小貨車首次通過該測站之交通量以及通過"03F0337S" 測站至"03F0394S" 測站的交通量

2.2.4 以 Accuracy 為準則訓練模型

將上面所提到的準則由 Specificity 轉為準確率，觀察模型表現好壞，其結果如下圖。

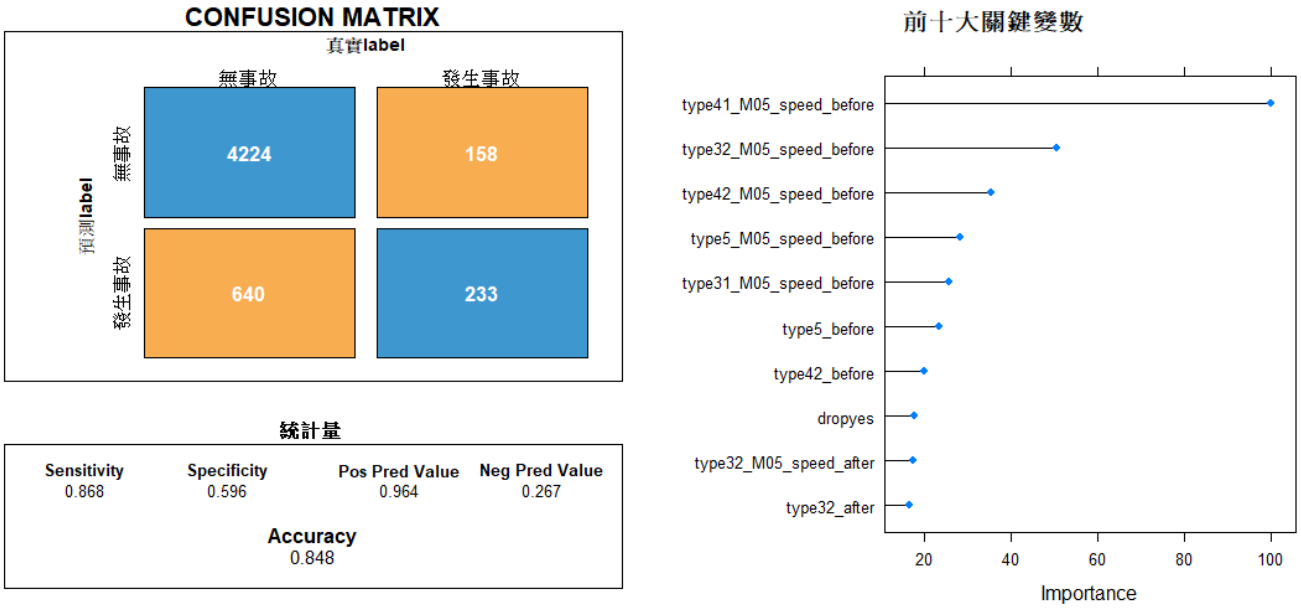
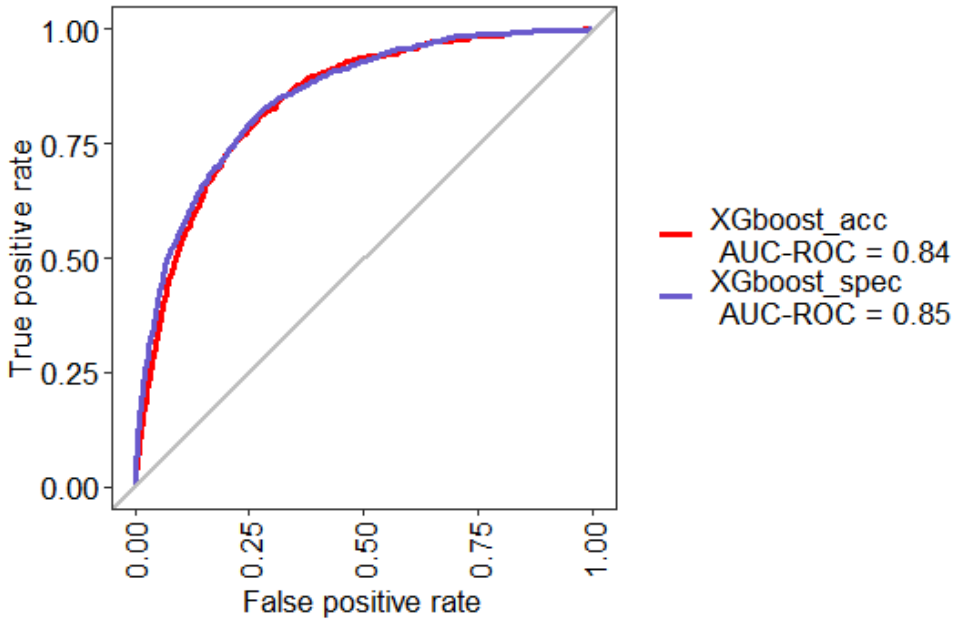


Figure 6: 混淆矩陣

可以看到 Specificity 由 0.749 下降至 0.596，PPV 下降至 0.964，然而準確率卻上升至 0.848。至於重要變數最主要的還是有各個車種的速度，其中較有差異的變化為 M07A 中的交通量資訊被凸顯出來。

2.2.5 兩模型之選擇

我們提供上面所配適的兩個模型之 ROC 曲線以及 AUC(於下頁)。
可以發現，其實兩個模型的 ROC 及 AUC 並未相差太多，若非得提供一個模型作為預報模型，我們將會提供以 Specificity 為準則的模型。雖然該模型犧牲了準確率。但是發生事故且未收到提醒，相較於未發生事故且收到提醒對於用路人的損失更大。接下來我們會介紹 SHAP value。



2.2.6 SHAP value

SHAP value 是某樣本內的每個變數對該樣本之預測數值貢獻，我們參考下方圖案[8]，可以看到同一樣本之不同變數會各自提供一個 SHAP value，將 SHAP value 加總可以得到一個 score，透過這個 score 能夠進行分類預測 (譬如：score 5 以上陽性等等)。現在我們專注於 SHAP value，倘若 SHAP value 的變異數越大，那該變數就會越顯得重要[9]。現在對事故資料配適以 Specificity 為準則的 XGBoost 模型後計算 SHAP value 並將變異數由大排至小提供於下方圖中，可以發現和上個章節我們辨識出的重要變數排序幾乎相同，”01F0339N” 至”01F0293N” 以及”01F0339S””01F0376S” 之間的小客車的車速為本預測之關鍵。

[8]假若某 A 同學年齡 18，BMI：17，血壓：120，性別為男，他得到的 score 為 3.5

[9]該數值變異數越大，也就較有機會出現極端數值。若該變數給該樣本之 score 一極大或極小的貢獻，其他變異較小的參數所給的貢獻就難以撼動整個 score 的變化。所以 SHAP value 變異越大的變數，對模型貢獻會更多，因為他主宰了 score 的大小。

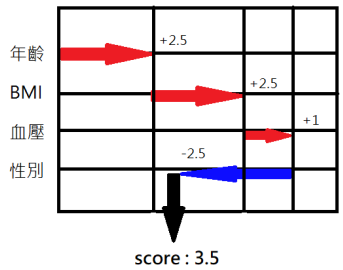


Figure 7: SHAP value 示意圖

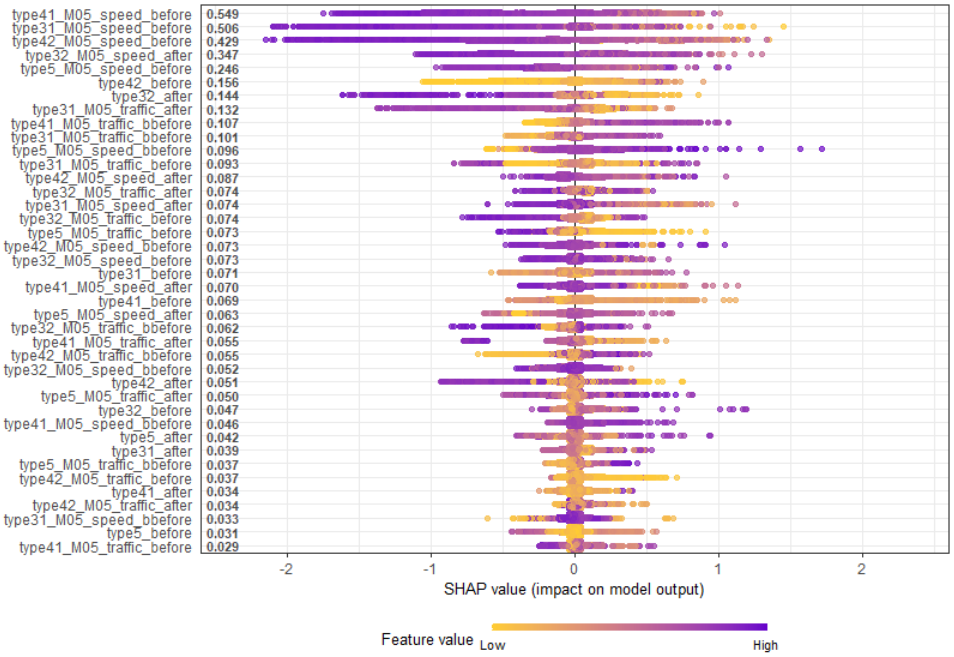
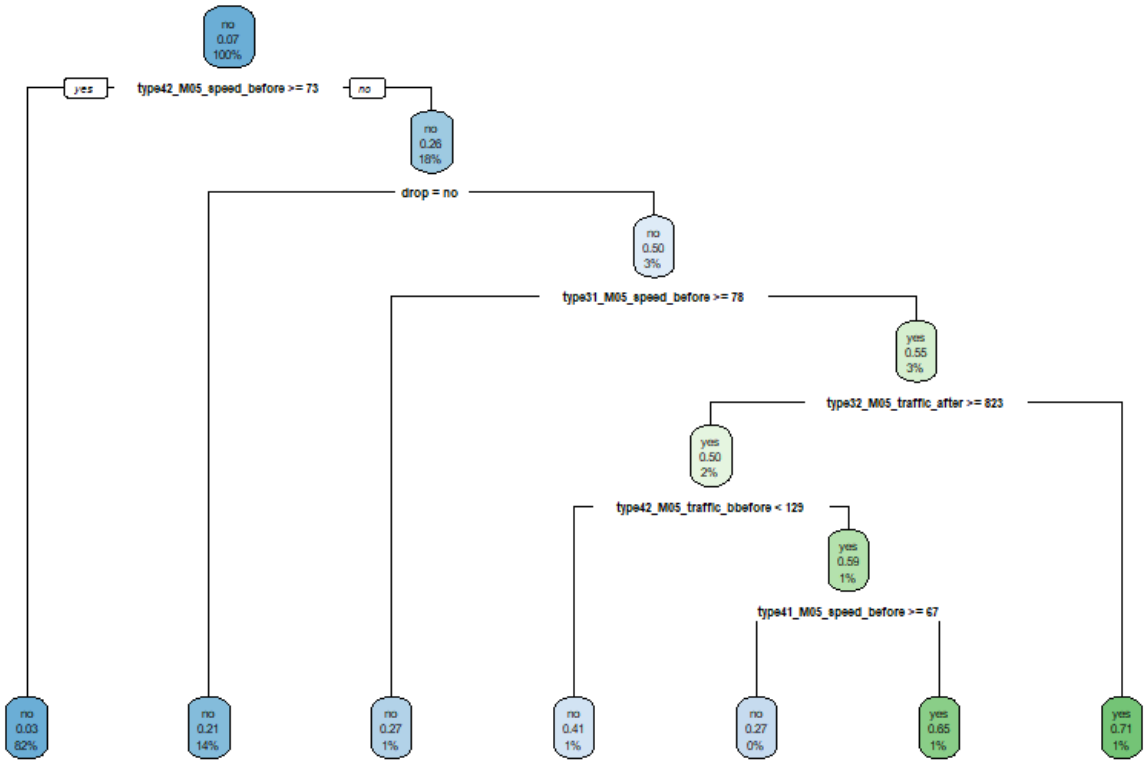


Figure 8: SHAP value

可以發現 SHAP value 點出之變數，與上方提供之重要變數相去無幾，頭五個都是和車速有關的變數。

2.2.7 國道事件與事故之關聯性探討

上面提到重要變數以及 SHAP value 兩個概念，雖然幫我們找出決定性的因素，但是並未告訴我們發生事故之情境為和，所以我們以 Specificity 為準則額外配適一個較好解釋的二元分類樹模型，配適完成後將二元樹架構畫出來，結果如下。



舉最右邊的路徑說明如何解釋二元樹，當”before” 路段之大貨車車速小於 73 公里且有掉落物，同時”before” 路段之小客車速度小於 78，恰好小貨車於”after” 路段車流小於 823，就容易發生事故 (此順序不可變更)。

總結來說，車速慢且車流量大，就存在一定的比重會發生事故。同時若相鄰路段間車速相差過大，由於車速瞬間遞減，用路人就容易發生事故。

3 結論以及後續努力方向

回顧前面章節，我們透過探索式資料分析找出國道一號 33.9 公里處有異常高的事故發生數量，接著針對此地區進行事故預測，將資料整理後以 XGBoost 模型建模，得到準確率、Specificity 以及 AUC 表現十分優異 (根據過往經驗，與人行為相關的資料做統計存在相當大之變動，能控制或是紀錄的變數較少)，且發現此區域車流速度慢時，易發生事故，若能夠適當提醒用路人注意行車安全，將會降低事故的發生。

另外能夠改進的地方有三點，第一點是預測間格為一小時。在收集資料的過程中，M05A 為每五分鐘更新一次資料，而 M07A 為一小時更新一次資料，並於隔天才提供資料，若我們要用這個模型進行" 即時" 預報，或許不能納入 M07A。雖然透過上面模型中所提供的重要變數可以發現，M07A 的資料看似未帶來多大的幫助，不過試圖把 M07A 資料所提供的變數移除後配適模型發現 Specificity 會下降，準確率也會下降，故還是建議納入 M07A 的資料。若能夠每隔五分鐘收集 M07A 的交通量變數，或許能夠配適出每五分鐘更新一次的預報模型，如此一來就稱得上是" 即時" 預報模型。

第二點是掉落物的時間界定問題，有些掉落物遺留在高速公路的時間橫跨多個小時，例如某物於 9：50 分掉落，10：20 分排除，很難界定存在掉落物的時間，所以存在一些偏差，關於時間界定還能夠更進一步探討。

第三點是變數的增減，我並非高速公路專家，或許有些變數應放入模型中而未被考慮，還需要與此領域之專家進行相關探討。

對前面提到剩下三個易肇事路段分別依照上述方式進行預測，就結果來說其準確率皆較"01F0339N/S" 地段高，而發生原因不盡相同。可見此模型泛用性高，下一步將可應用於每個測站，若有興趣我們能再做後續的努力，謝謝觀看。

4 參考資料

1. 行政院公布之中華民國 108 年（西元 2019 年）政府行政機關辦公日曆表
   (https://www.dgpa.gov.tw/information?uid=83&pid=8150)

2. 交通部高公局公布之交流道、服務區里程一覽表
   (https://www.freeway.gov.tw/Publish.aspx?cnid=1906)

3. 交通部高公局公布之交通資料庫
   (http://tisvcloud.freeway.gov.tw/history/)

4. 交通部高公局之交通安全資訊
   (https://www.freeway.gov.tw/Publish.aspx?cnid=516&p=2849)

5. python scikit learn Machine Learning in Python
   (https://scikitlearn.org/stable/modules/cross_validation.html)

6. Research Gate
   (https://www.researchgate.net/figure/Calculationofsensitivity-specificityandpositiveandnegativepredictive_fig1_49650721)