

MDS Final Project

道路事故嚴重度預測與風險熱點預測

Group L - 施丞澤、江彥宏、陳奕廷、廖振翔



目錄




- 研究動機與議題
- 資料集分析
- 嚴重度重要影響因子分析
- 模型訓練與預測
- 時空風險預測模型
- Demo



研究動機與議題



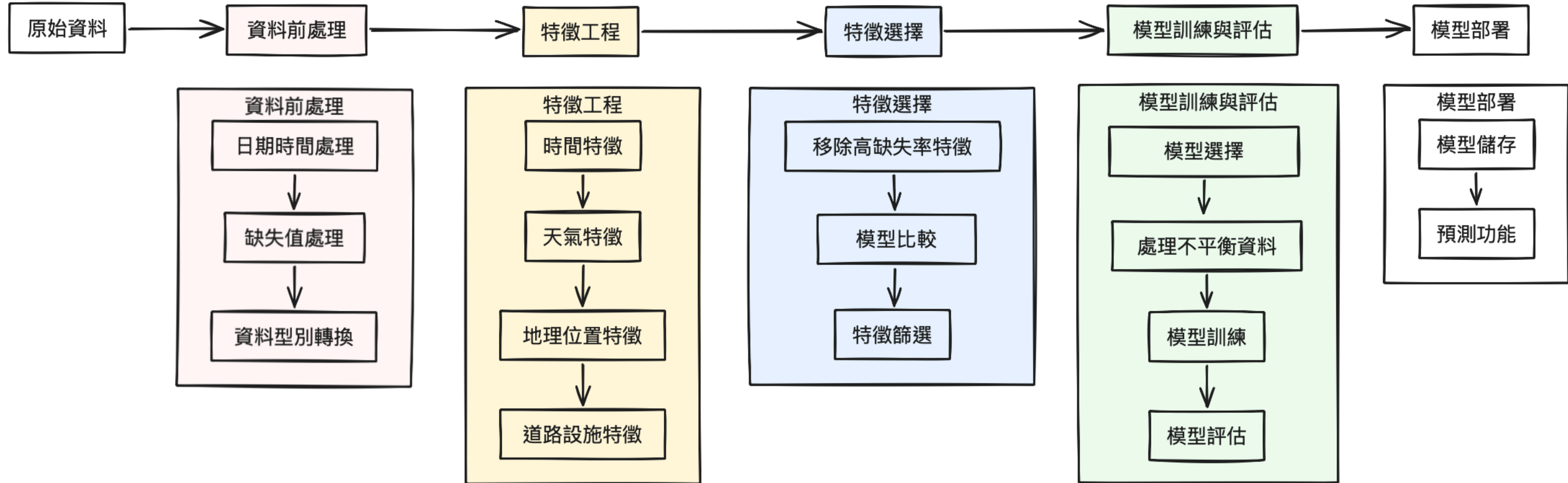
研究動機

- 
- 美國 2016-2023 期間 累計 7.7M 筆事故記錄，嚴重事故造成逾百億美元損失。
 - 利用 US Accidents 龐大事故資料，建立時空風險預測模型
 - 協助警力與資源精準部署，降低成本、事故與傷亡

目標展開議題樹



Flow Chart





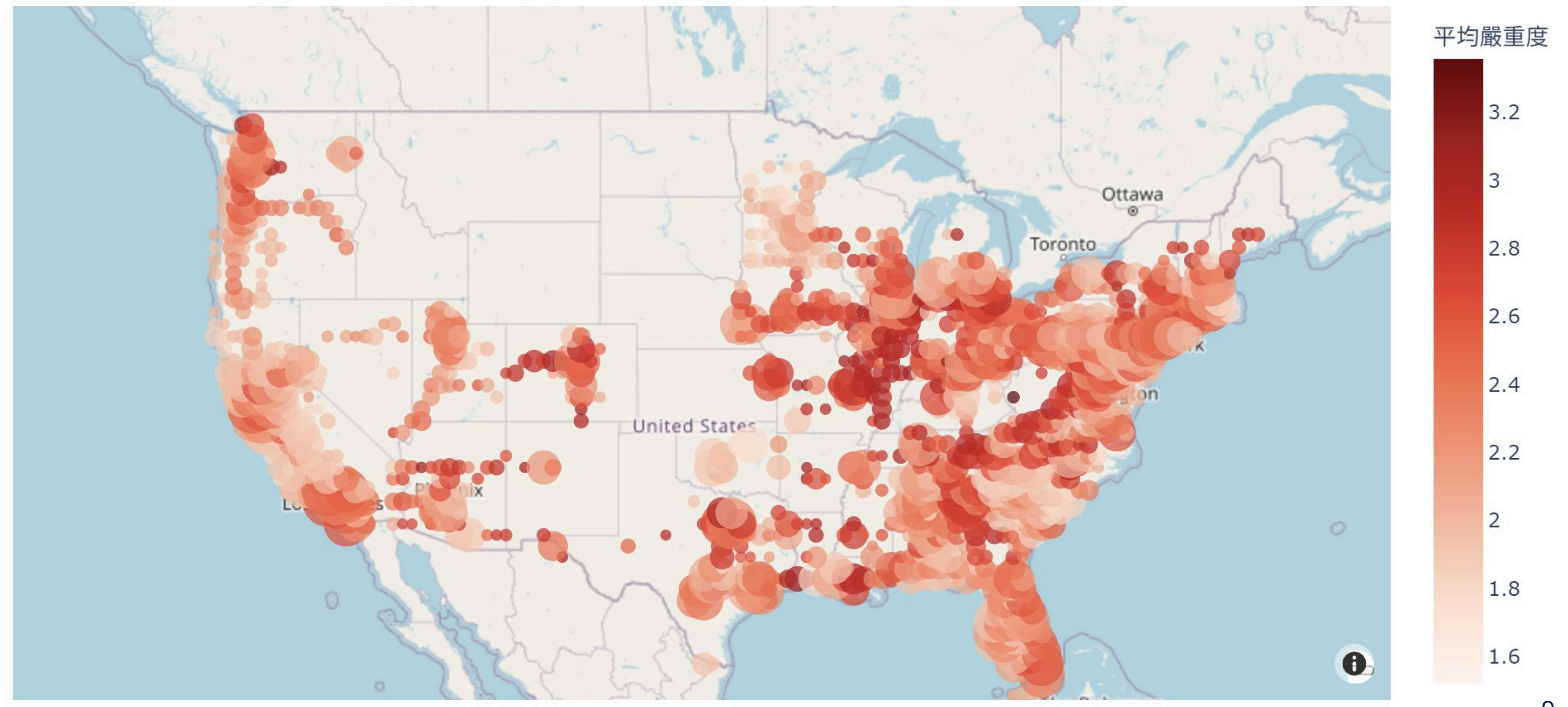
資料集分析



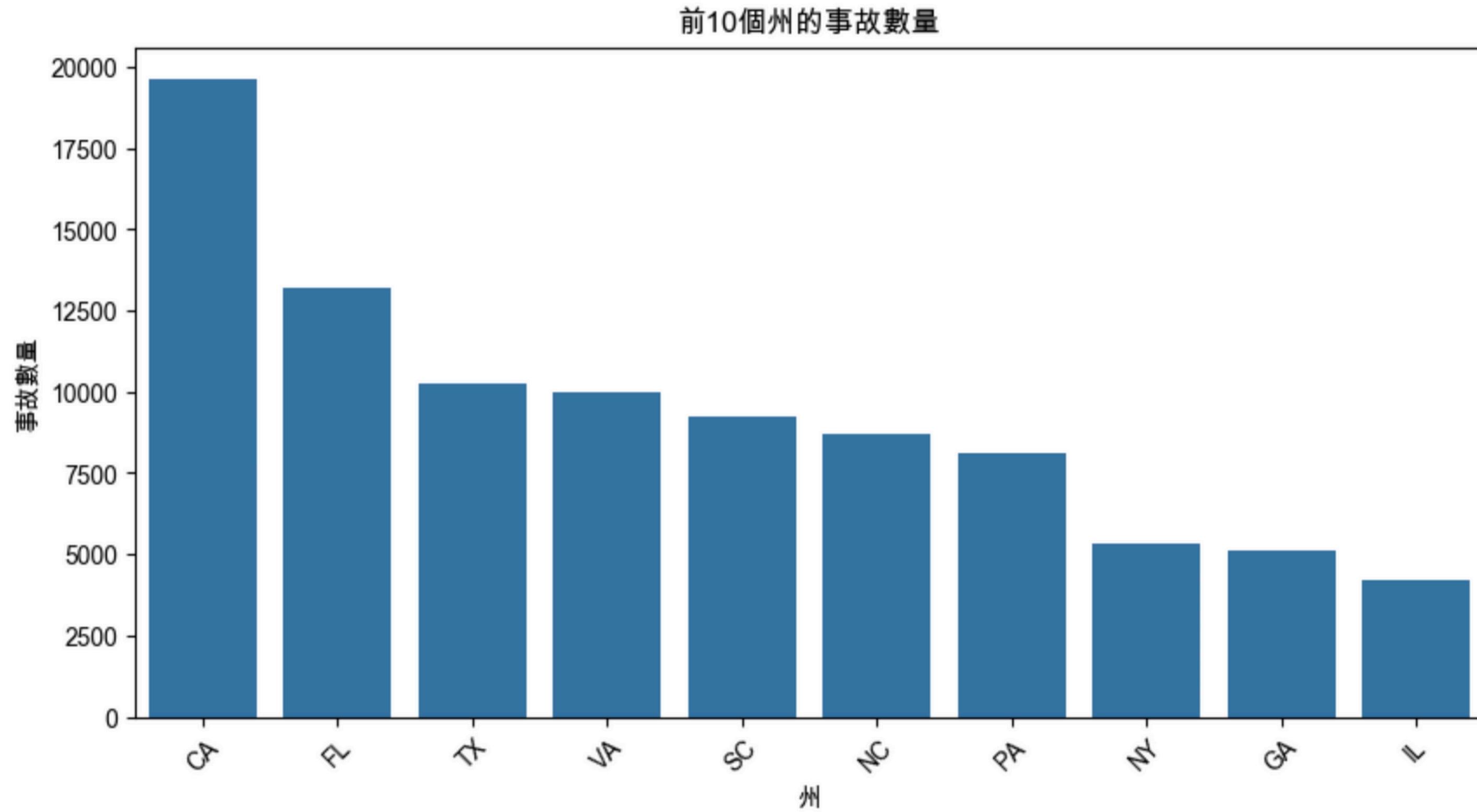
資料集說明

欄位名稱	說明
ID	事故記錄的唯一識別碼
Source	原始事故資料的來源
Severity 	事故嚴重程度，數字介於 1 到 4，1 表示對交通影響最小（短暫）
Start_Time	事故發生的開始時間（當地時區）
End_Time	事故影響結束的時間（當地時區）
Start_Lat, Start_Lng	事故起點的 GPS 緯度和經度
End_Lat, End_Lng	事故終點的 GPS 緯度和經度
Distance(mi)	事故影響路段的長度（英里）
Description	事故的文字描述
Street, City, County, State, Zipcode, Country	事故發生地的地址資訊
Timezone	事故地點所屬的時區（如東部時間、中部時間等）

資料集分析 - 美國交通事故熱力圖



資料集分析 - 前十個州的事故數量





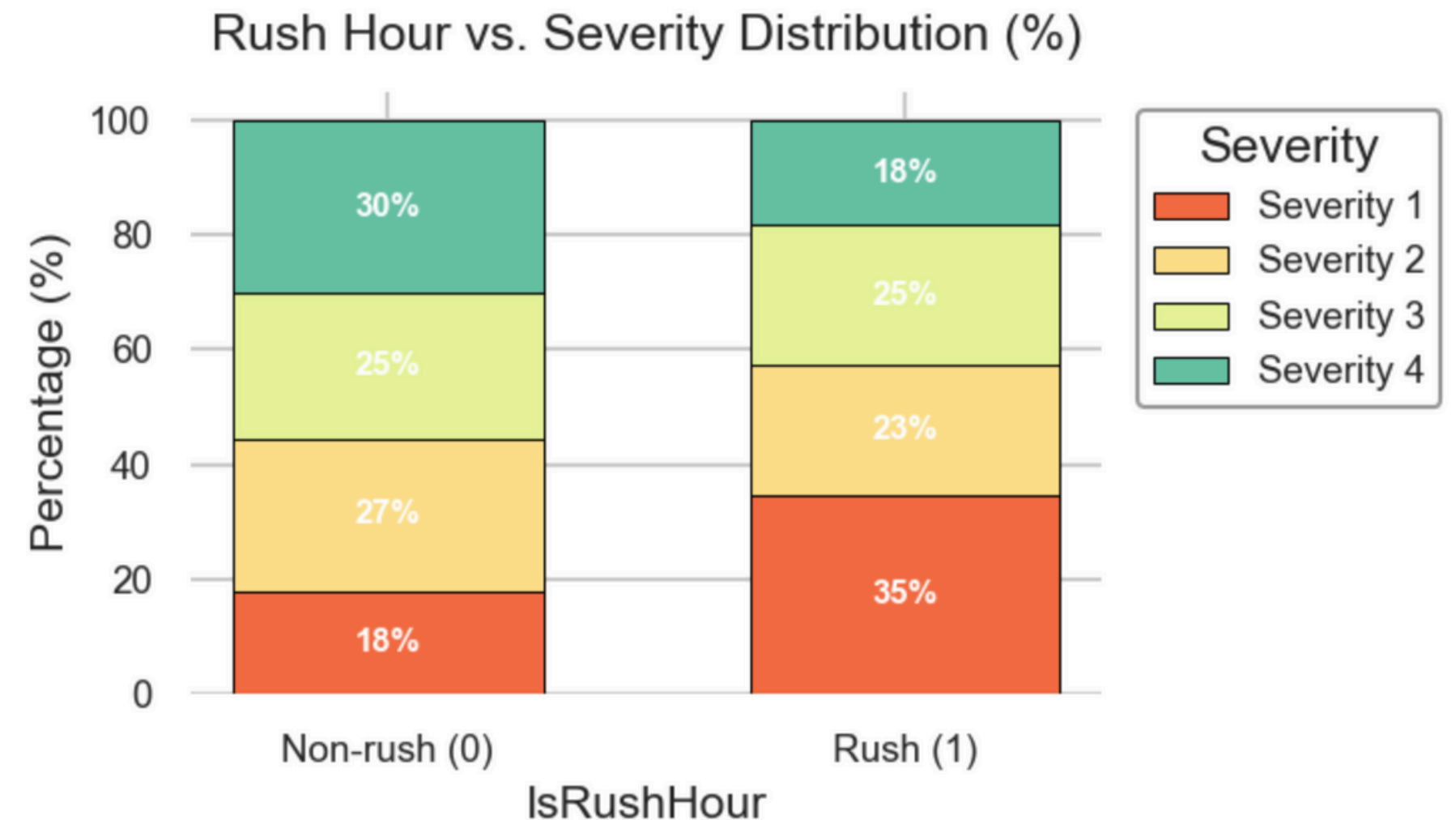
嚴重度重要影響因子分析



Feature Engineering

- 時間特徵衍生：

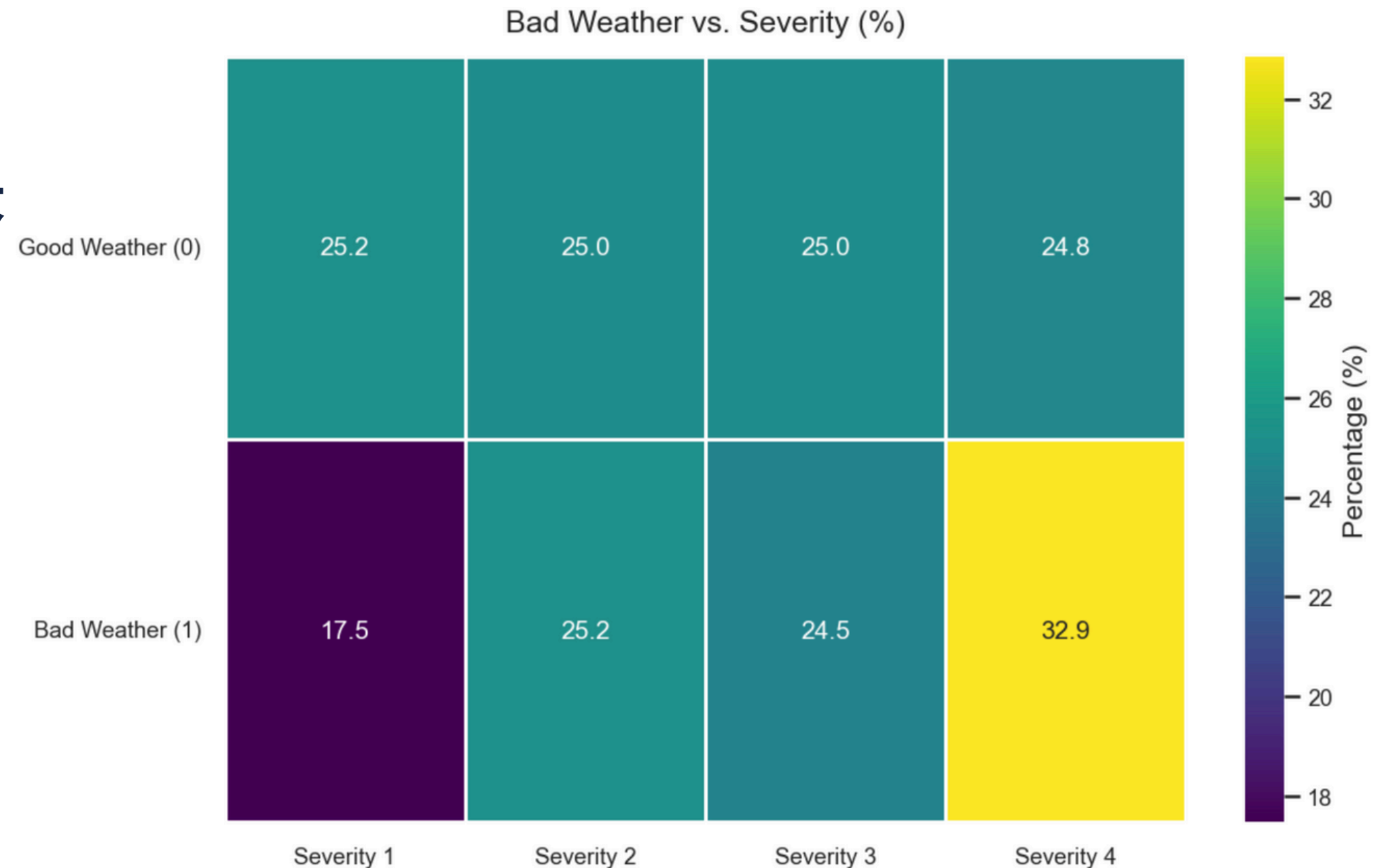
- 直接從「**Start_Time**」抽取小時、星期、月份，並進一步轉成**尖峰／深夜／週末**等二元標籤
- 補充週期性編碼 (**Sin／Cos**) 讓模型理解 **23 點接近 0 點、星期六接近星期日、月末接近月初**等週期關係



Feature Engineering

- 天氣類別標準化：

- 將原本過多的**天氣文字**（如 “Light Rain”、“Mostly Cloudy”）歸類為 **7 大天氣類別**（Clear、Cloudy、Rain、Snow、Fog、Windy、Other），減少類別稀疏
- 衍生「IsBadWeather」「IsGoodVisibility」「IsFreezingPrecip」等二元標籤，強化極端天候情境的風險訊號



Feature Engineering

- **地理／道路特徵擴充：**
 - 判定「IsUrban」：有無重要路口、行人道、交通號誌等，標示地點**是否屬於都市化區域**
 - 計算「Road_Complexity」：該地點擁有幾種不同道路設施，量化**交通環境複雜程度**
 - 加入交互作用：「Signal_x_BadWeather」，揭示**惡劣天氣下交通號誌存在時風險是否放大**

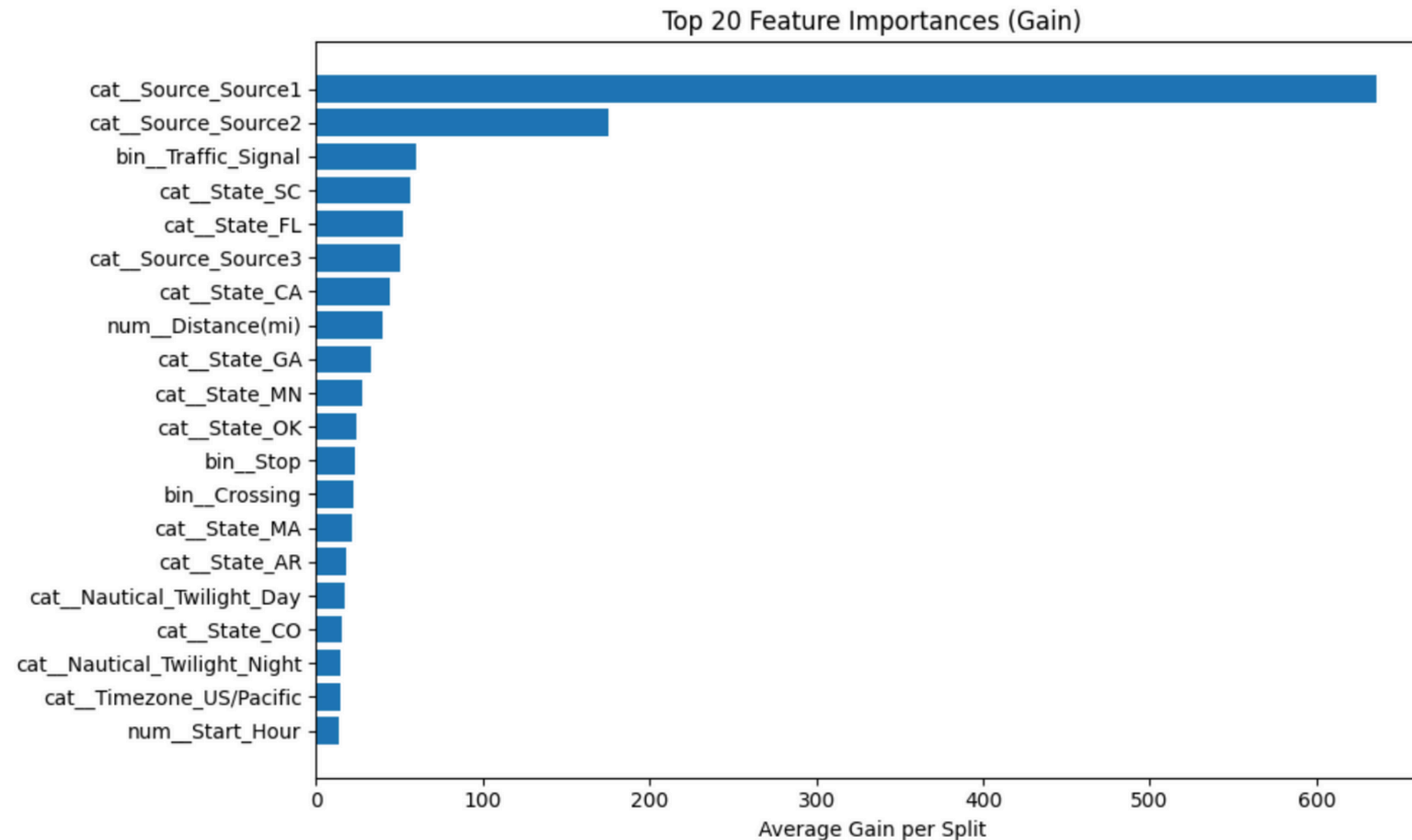
Feature Engineering

- 後續流程

- 分割資料並建立 Pipeline（數值直通＋類別 One-Hot）
- 模型訓練與特徵選擇，篩出最具預測力的子集
- 可視化檢視，驗證各特徵與嚴重度關係
- 迭代優化：針對無貢獻或過度相關的特徵進行剔除或重設計

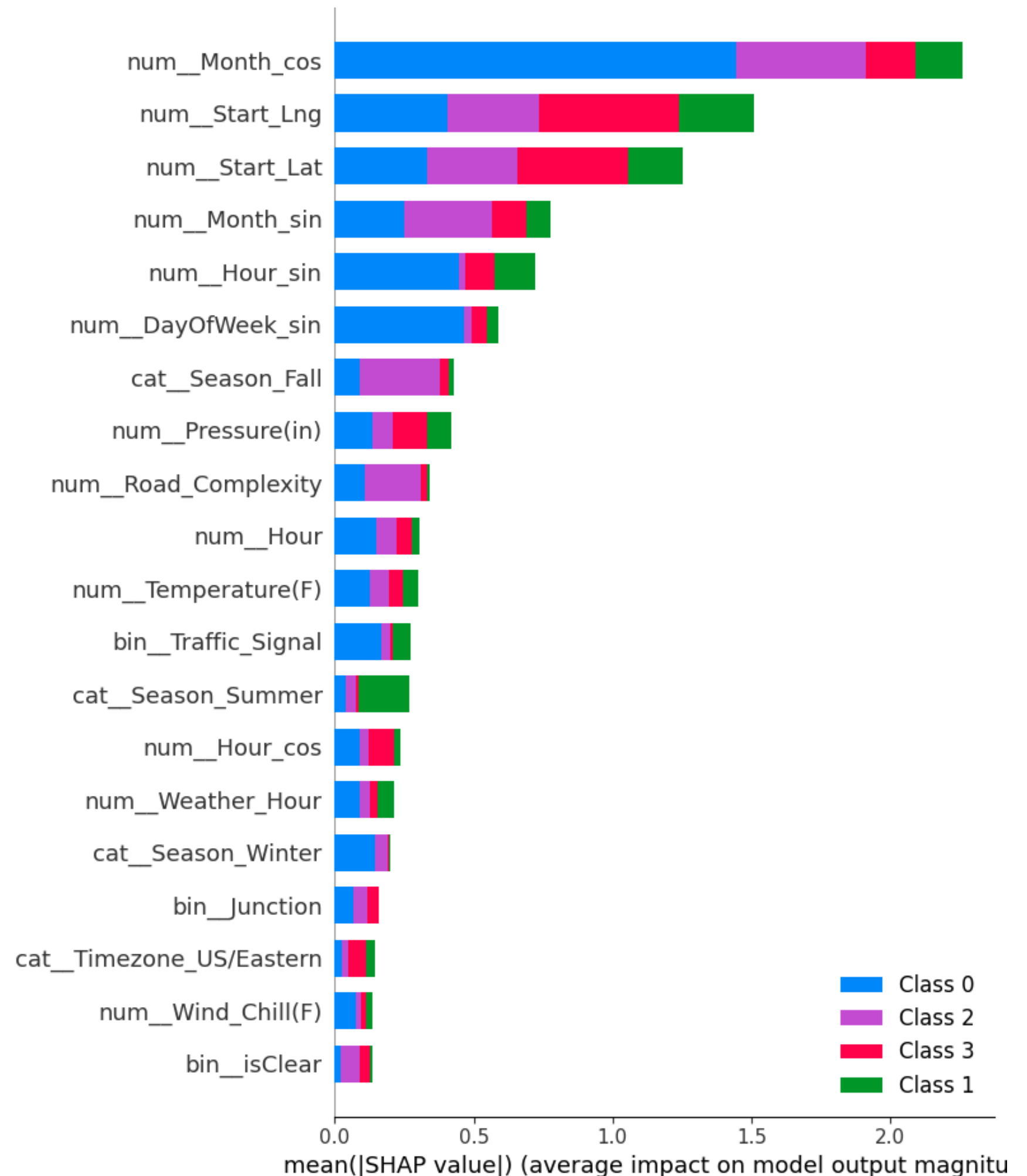
Feature Selection

- 迭代發現：許多特徵有偷看答案的嫌疑，觀察 F1-score 變化，例如：Source



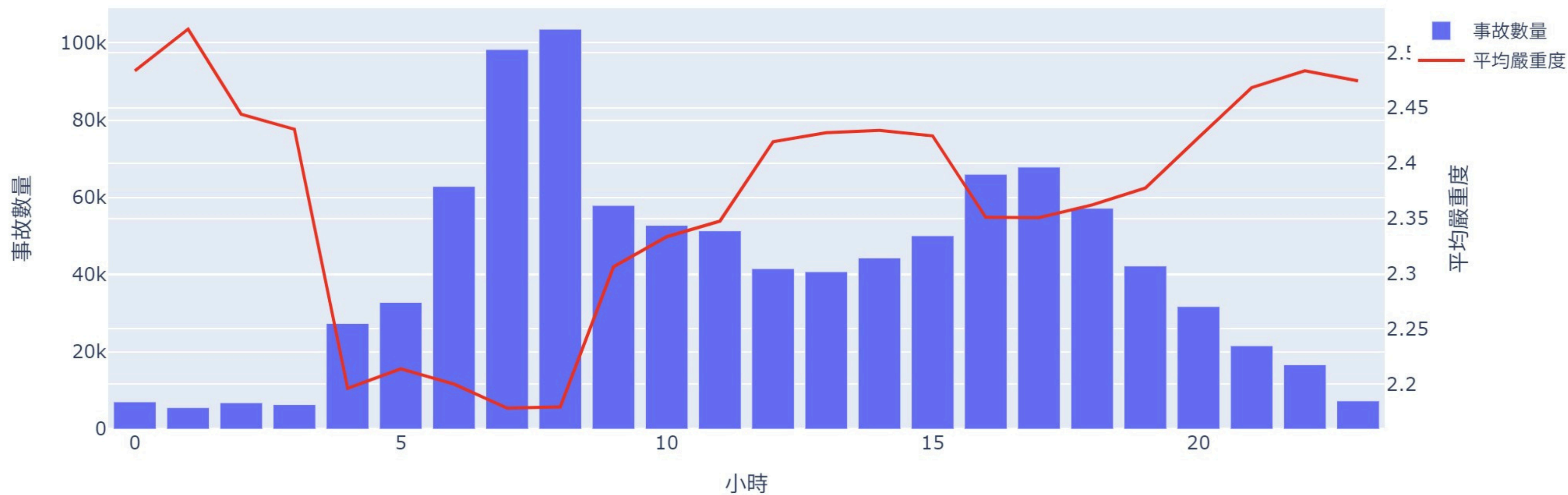
Feature Selection

- 使用 SHAP 特徵影響力分析
- 時間相關特徵：
 - 月份、星期、季節、小時
- 環境相關特徵：
 - 溫度、氣溫、風寒指數
- 位置相關特徵：
 - 經緯度

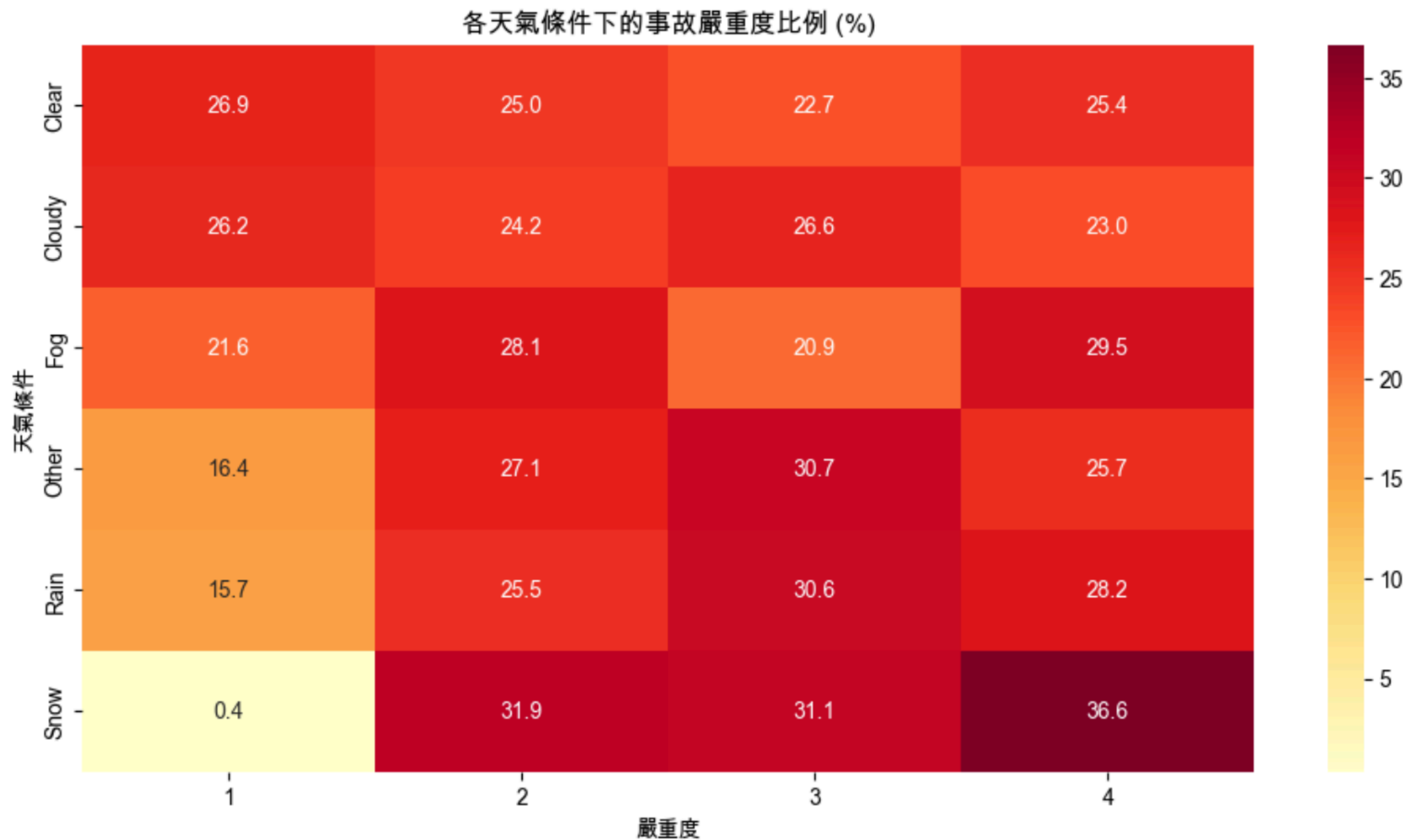


影響因子解釋 - 當日時間點

24小時事故分布



影響因子解釋 - 天氣狀況





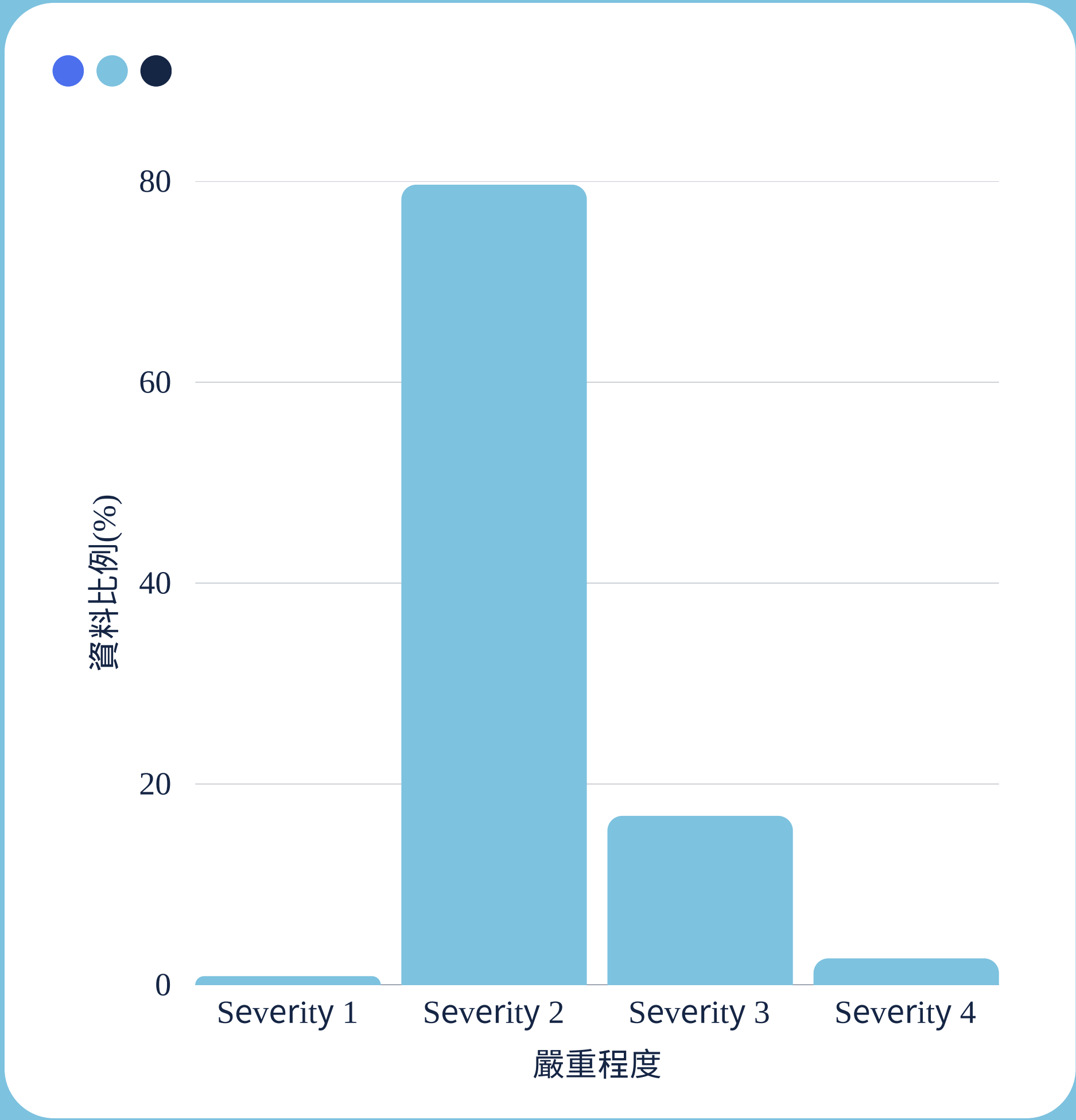
模型選擇與訓練



資料拆分與預測目標

- 目標：預測事故嚴重度 (Severity) (1-4)
- 資料拆分：
 - 2016-01 至 2022-03 (訓練集)
 - 2022-04 至 2023-03 (測試集)
 - 用時間切分而非隨機抽樣，避免把未來資訊洩漏到過去模型。
 - 以整年資料做為測試集，更能代表整體週期變化。
 - 訓練共 6,177,254 筆資料，測試共 1,544,314 筆。

事故嚴重度 類別比例





資料不平衡處理

- 如果只看普通 Accuracy，模型光是全猜 Severity 2 就有近八成命中
- 採用 Kaggle 上參考的最佳採樣策略：
 - 混合採樣 (Mixed Sampling)：讓所有的類別的樣本數相近
- 採樣後樣本數：
 - 類別 1: 53,892 \rightarrow 901,654
 - 類別 2: 4,921,156 \rightarrow 901,654
 - 類別 3: 1,039,401 \rightarrow 901,654
 - 類別 4: 162,805 \rightarrow 901,654

模型選擇與評估


- 使用 Kaggle 常見 TOP 5 模型：
 - LightGBM
 - XGBoost
 - CatBoost
 - Balanced Random Forest
 - Neural Network
- Severity 2 佔 79%+，傳統 Accuracy 易被多數類別主導。
 - 使用 Balance Accuracy 評估 → 反映少數類預測能力。

模型比較

model	Accuracy	F1-score	Balance Accuracy	Training Time(s)
LightGBM	0.6804	0.7300	0.7213	297.51
 XGBoost	0.6543	0.7090	 0.7494	37.76
CatBoost	0.6189	0.6791	0.7167	45.96
Balanced Random Forest	0.5924	0.6567	0.6953	135.86
Neural Network	0.0545	0.0350	0.5035	1089.64

有無特徵工程/採樣比較

評分標準：Balance Accuracy

Feature Engineering	Mixed Sampling	LightGBM	 XGBoost	CatBoost
		0.4740	0.4979	0.4323
	✓	0.7246	0.7320	0.7146
✓		0.4725	0.5006	0.4325
✓	✓	0.7262	0.7362	0.7155

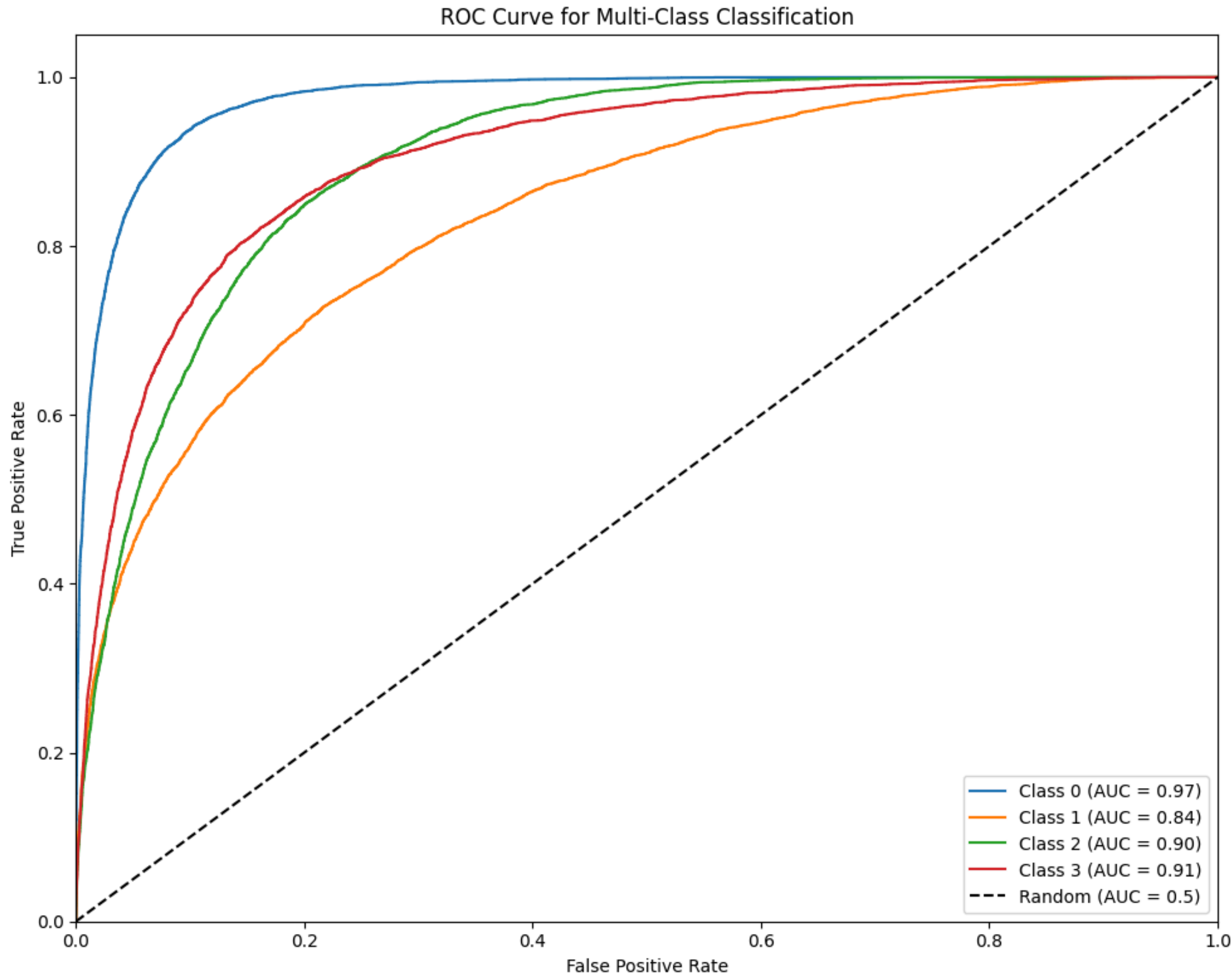
XGBoost 訓練結果

	Precision	Recall	F1-score
Severity 1	0.8011	0.8997	0.8475
Severity 2	0.6754	0.6149	0.6443
Severity 3	0.6457	0.7157	0.6789
Severity 4	0.7053	0.7096	0.7074

- 最終訓練結果之平衡準確率 (Balance Accurary) 為：**0.7349**

XGBoost 訓練結果

	Per-class ROC-AUC
Severity 1	0.97
Severity 2	0.84
Severity 3	0.91
Severity 4	0.91





時空風險預測模型



時空風險預測模型

- 考慮預測資料的大小 → 只針對 **California(CA)**進行預測
 - CA 事故量佔全美 18%
 - 預測時間尺度為 2023-01 至 2023-03
- 嚴重度 (Severity) → 加權嚴重度 (Weighted Severity)
 - XGBoost 對每筆事故輸出 4 維機率向量 $P = (p_1, p_2, p_3, p_4)$, 為 Severity 1, 2, 3,4 的機率
 - **Weighted Severity** = $\sum p_i \cdot \text{Severity}_i$
 - 把多類別預測的**完整不確定性**壓縮成一個**連續值** (1~4) 。
 - Kepler 熱度圖：> 3.0 = 高風險路段



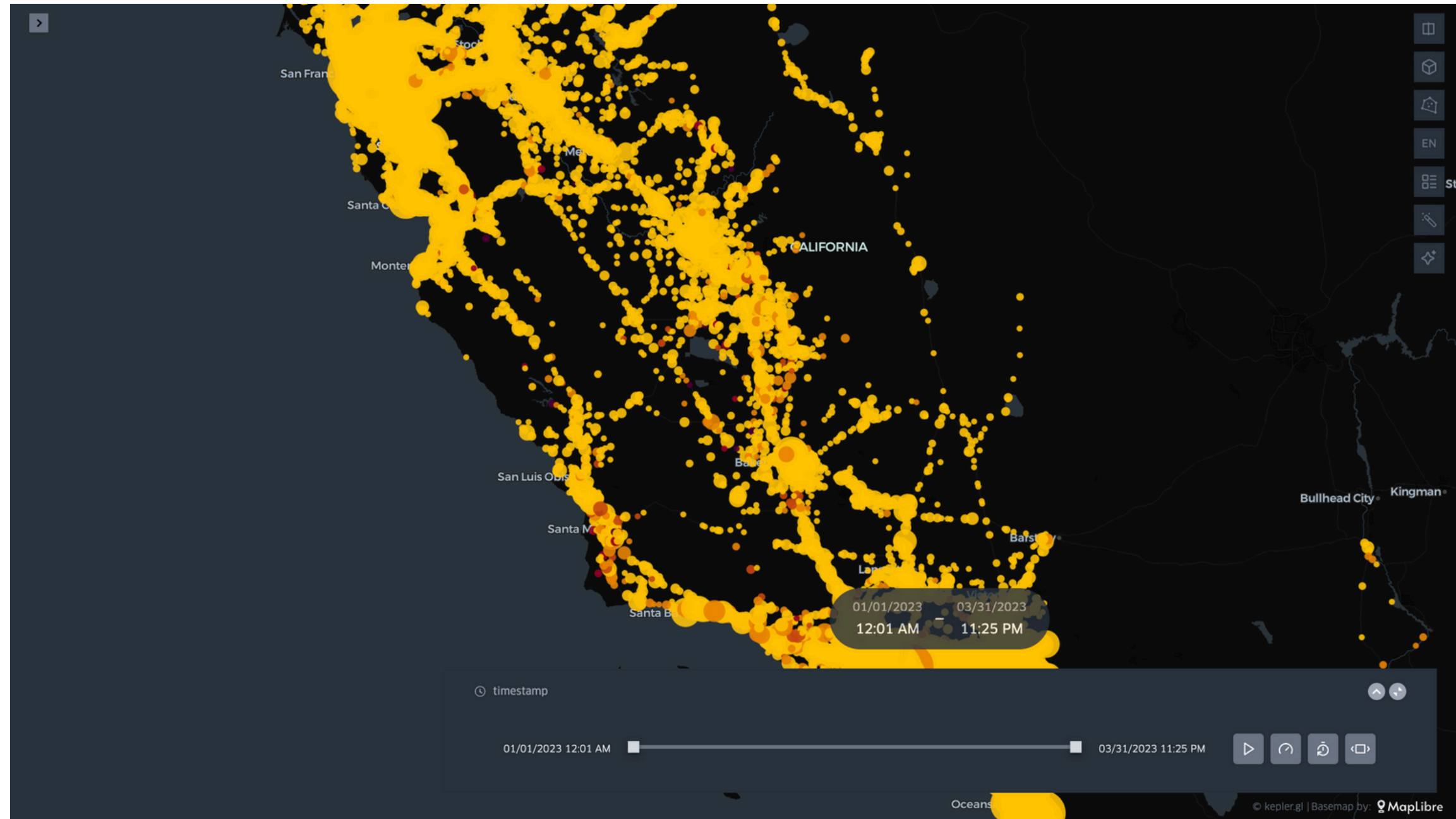
Demo 影片



Demo 系統介紹 (kepler.gl)

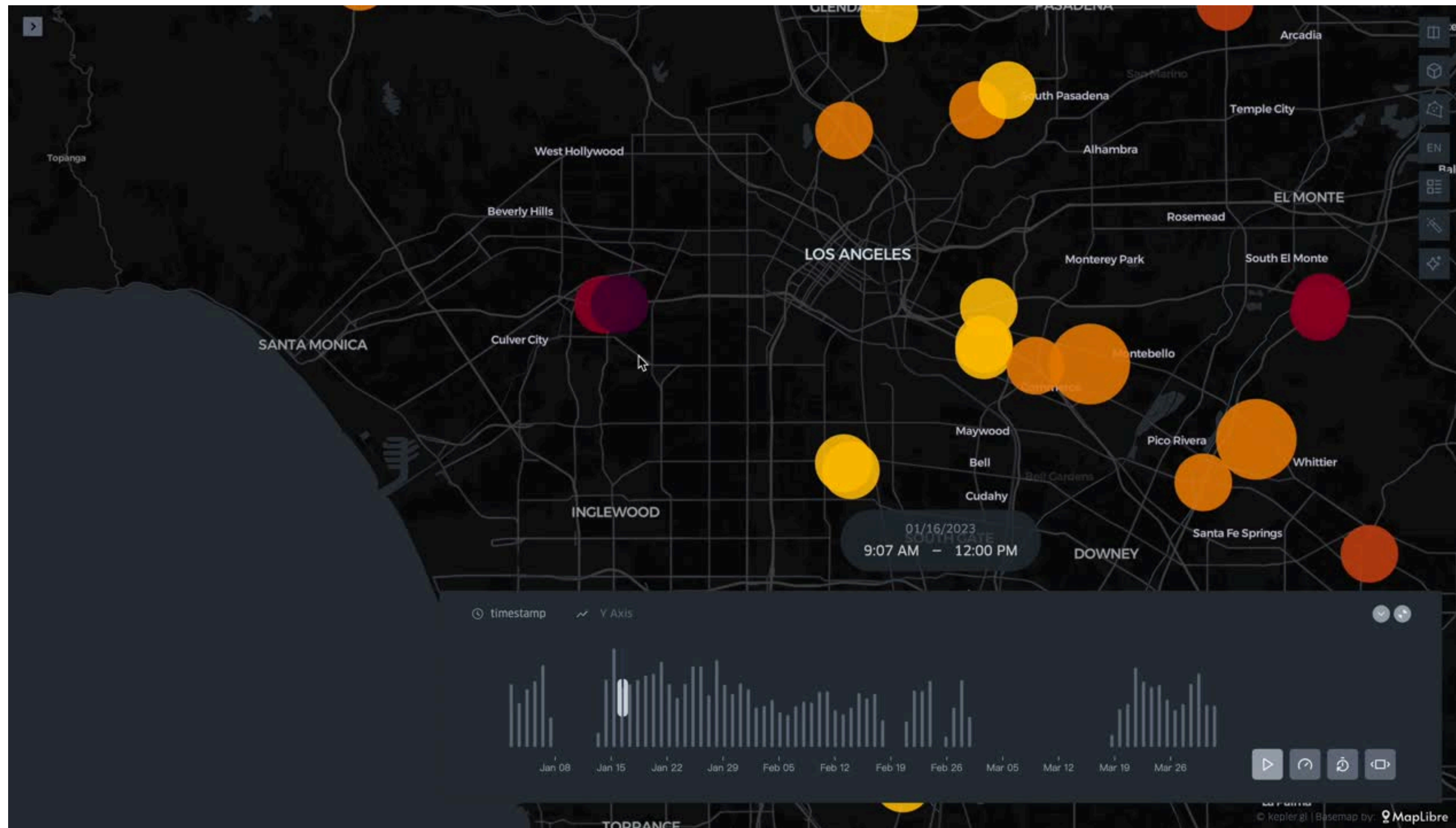
展示車禍發生地點與預測嚴重性

- **Time Playback:** 展示不同時間點的車禍情況
- **Filter:** 篩選不同天氣條件下的車禍情形
- 預測未來時間點可能發生車禍分布之嚴重程度
- 只展示加州的車禍分佈




以 2016-2022 年資料預測 2023/1 ~ 2023/3 間的交通事故情形


Demo 影片



應用策略

- 
- 即時監控與預警：針對高風險時段／路段發佈警示。
 - 資源優化配置：依模型預測調度警力與救護。
 - 基建與政策調整：根據嚴重度趨勢優化路面與宣導。

未來展望

- 
- 風險清單：不只視覺化，能夠具體輸出風險路段。
 - 真實天氣資料：讓模型可以採用真實的天氣預測資料，評估風險。
 - 擴展地區：擴展預測模型的範圍。



謝謝大家！