SINGAPORE 2019

26th ITS World Congress
21–25 October

Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting

Mihaita, A.S., Liu, Z., Cai, C., Rizoiu, M.A

Speaker Name

Adriana-Simona Mihaita
University of Technology Sydney, Australia

Summary

Introduction

1. Almost 60% of traffic congestion is due to non-recurrent incidents [1]
2. Australia - the annual economic cost of road crashes ≈ $27 billion/y (2017) [2]
3. Various factors can influence the duration of traffic disruptions: location, time of day, severity, fatalities, proximity to public facilities, weather, etc.
4. Traditional methods for incident duration prediction:
   - linear/non-parametric regression models [4],
   - Bayesian classifiers [5],
   - discrete choice models (DCM) [6],
   - probabilistic distribution analyses [7], and
   - the hazard-based duration models (HBDM) [8]
5. Majority of work – undergone for motorway incident prediction/

[1] D. a. L. Schrank, T., "The annual urban mobility report," College Station, TX: Texas Transportation Institute2003
[2] A. Government. (2017, 25/07/2018). *Road Safety*. Available: https://infrastructure.gov.au/roads/safety/
[4] S. a. R. Peeta, J. and Gedela, S., "Providing real-time traffic advisory and route guidance to manage borman incidents on-line using the hoosier helper program " Purdue University, School of Civil Engineering2000.
[5] F. D. Boyles S, Waller ST (2007), "A Naive Bayesian Classifier for Incident Duration Prediction," presented at the TRB 86th Annual Meeting, Washington DC, United States., 2007.
[6] L. Ruimin, Z. Xiaoqiang, Y. Xinxin, C. Nan, and Z. Jianan, "Incident Duration Model on Urban Freeways Based on Discrete Choice Model," in *2010 International Conference on Electrical and Control Engineering*, 2010, pp. 3826-3829.
[7] G. Giuliano, "Incident characteristics, frequency, and duration on a high volume urban freeway," *Transportation Research Part A: General,* vol. 23, no. 5, pp. 387-396, 1989/09/01/ 1989.

Challenges

- How to accurately predict incident duration on arterial roads and regular streets in a large city?

- How to integrate traffic flow information in the prediction process? What road sections and from what timespan (before/after the incident was reported)?

- What would be the most influential factors which affect the incident duration that traffic centres need to prioritise for a fast and efficient incident clearance ?

Victoria Road Incident Duration Prediction – Sydney, AU

3. Data sources

- one year (2017) of traffic incidents reported by the Traffic Management Centre (TMC) in Sydney.
- contains 5,134 records of various planned and unplanned incidents:
  - hazards,
  - road closures to
  - accidents and
  - maintenance work.
- we focus on incidents labelled as "**Accidents**" – they induce the longest clearance time in the current subnetwork

- Total = **574** accident records
- mean duration of **44.59** minutes, a
- maximum duration of 719 minutes.

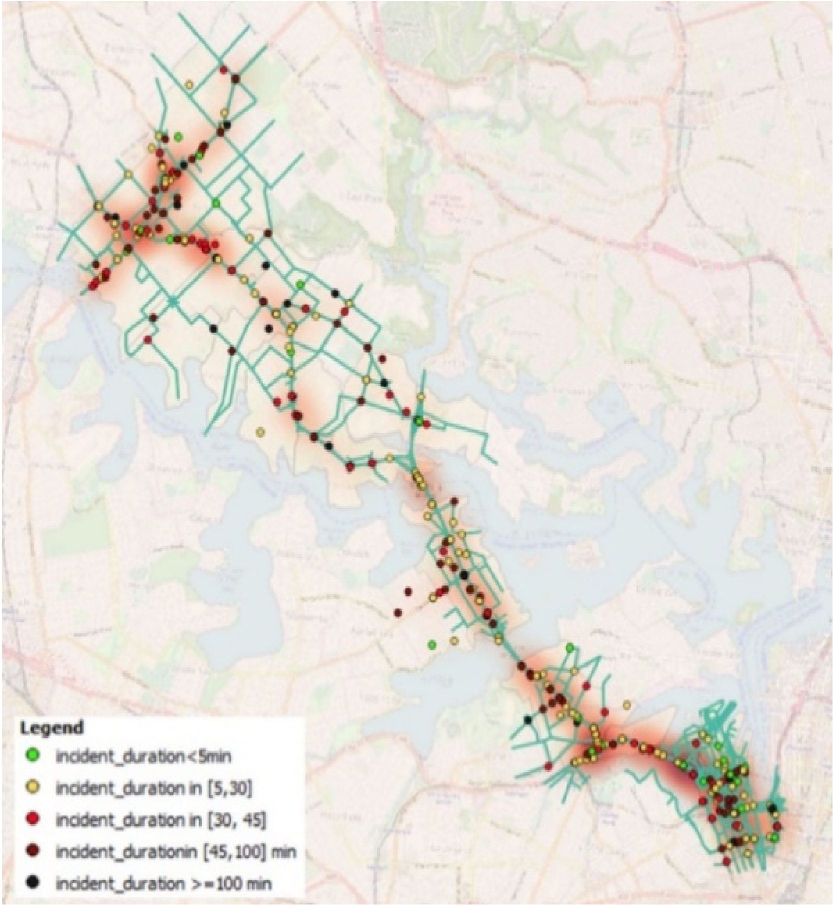


Fig 1. Heat-map the "accident duration" distribution in the Victoria Rd subnetwork.

Victoria Road – Sydney, AU

3. Data sources

- one year (2017) of traffic incidents reported by the Traffic Management Centre (TMC) in Sydney.
- contains 5,134 records of various planned and unplanned incidents:
    - hazards,
    - road closures to
    - accidents and
    - maintenance work.
- we focus on incidents labelled as "**Accidents**" – they induce the longest clearance time in the current subnetwork

- Total = **574** accident records
- mean duration of **44.59** minutes, a
- maximum duration of **719** minutes (12h).

Categories	*Features/explanation*	*Value dataset*
Accident	Location	{X,Y} in GDA Lambert coordinates
	Hour of day	{0,1,…23}
	Peak Hour	{1,0}
	Day of week	{1,..5}
	Weekend	{0,1}
	Month of the Year	{1,2,..12}
	Type	{Accident}
	Subtype	{Bus, car, bicycle, animals, etc.}
	Affected lanes	{Null, 1 lane, 2 lanes, 3 lanes, 4 lanes, All lanes, breakdown}
	Direction	{East, West, North, South, E-W, N-S, One Direction, Both Directions}
	Severity	{1,2,..10}
	Incident Source	{1,2,3}
	Unplanned	{0-planned,1-unplanned}
Weather	Average Temperature	ranging from {11.13 °C – 32.4 °C }
	Rainfall	ranging from {0 – 85mm}
Events	Public holidays	{0-no,1-yes}
Area geometry	Sector ID	As defined by TMC
	TZName	As defined by BTS (Bureau of Transport Statistics)
	Section ID	$\mathbb{R}_+$
	Section Speed	$\mathbb{R}_+$ [Km/h]
	Section Lanes	{0,1,2,3,4,5,6}
	Section Capacity	{0, max 3100 vehicles/hour}
	Section class	As defined by TMC
	Street ID	As defined by TMC
	Intersection ID	As defined by TMC
	Distance from CBD	$\mathbb{R}_+$ [Km]

Victoria Road – Sydney, AU

3.1 Data sources

- majority of accidents are cleared off in less than 30 min (291 out of 574)
- incident duration is long-tail distributed, with the longest 10% of the incidents (57 out of 574) spanning between 100 and 719 minutes
- ECCDF - incident duration presents two different regimes given by different slopes to the right and left of a threshold T (identified to be revolving around 45minutes)

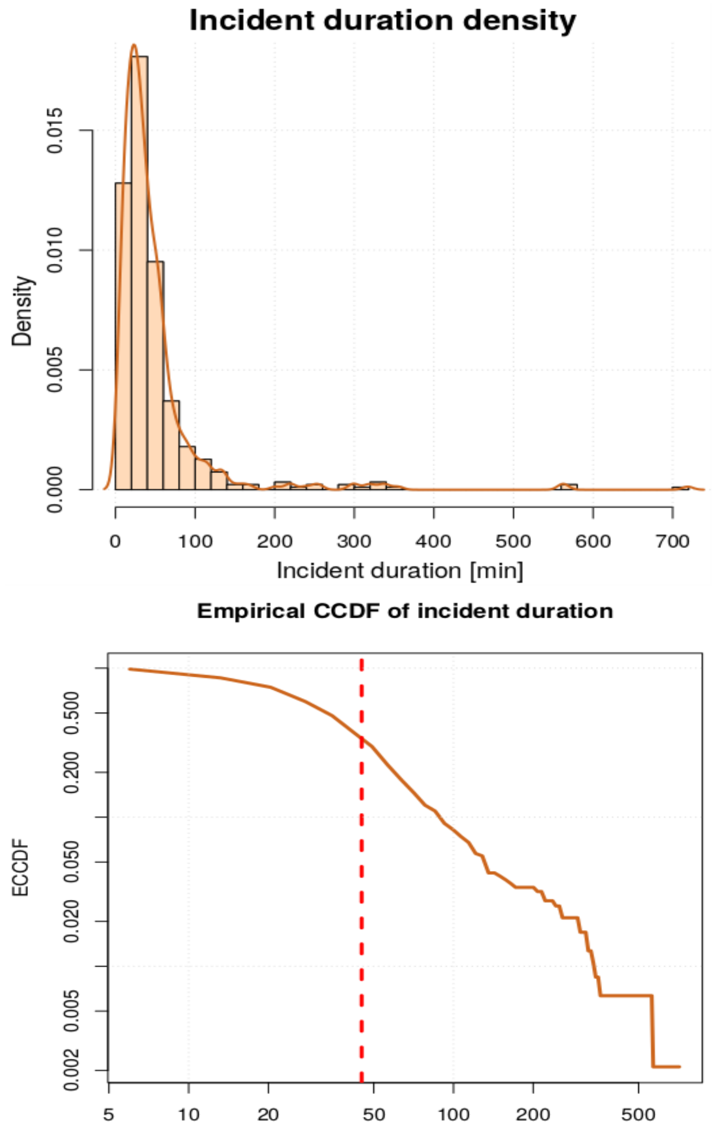
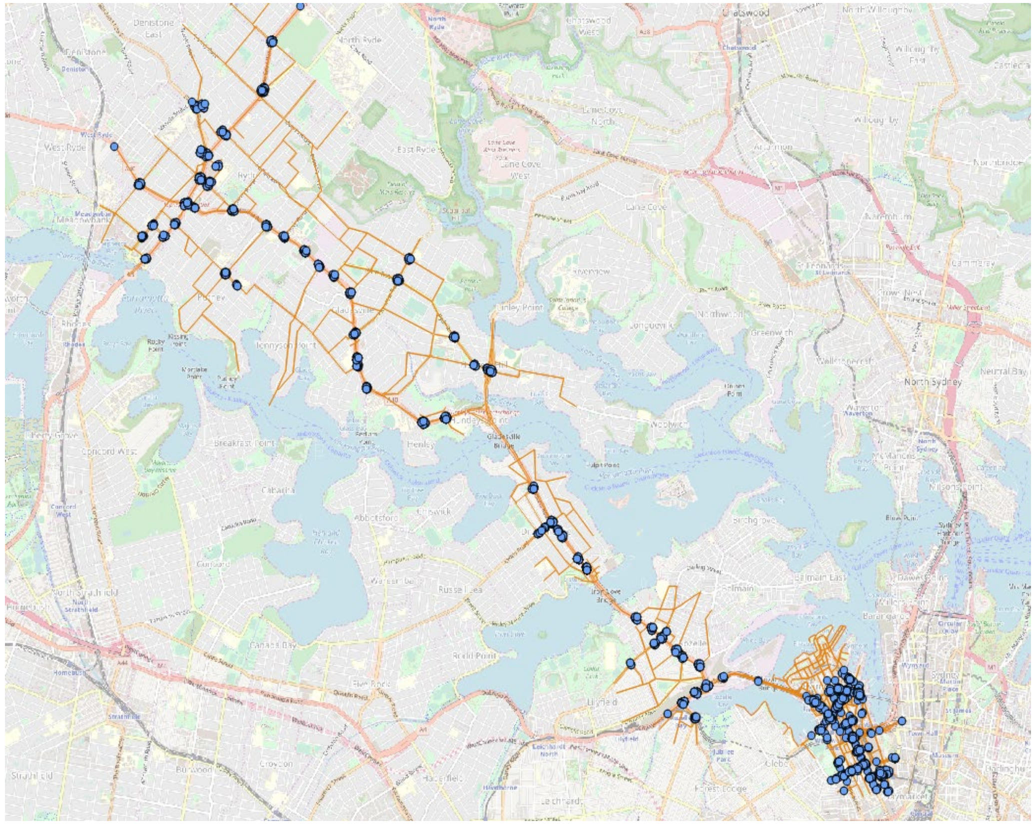


Fig 2
a) Accident duration distribution and
b) Empirical Cumulative Distribution Function of the accident duration.

Victoria Road – Sydney, AU

3.2 Adding in traffic flow features

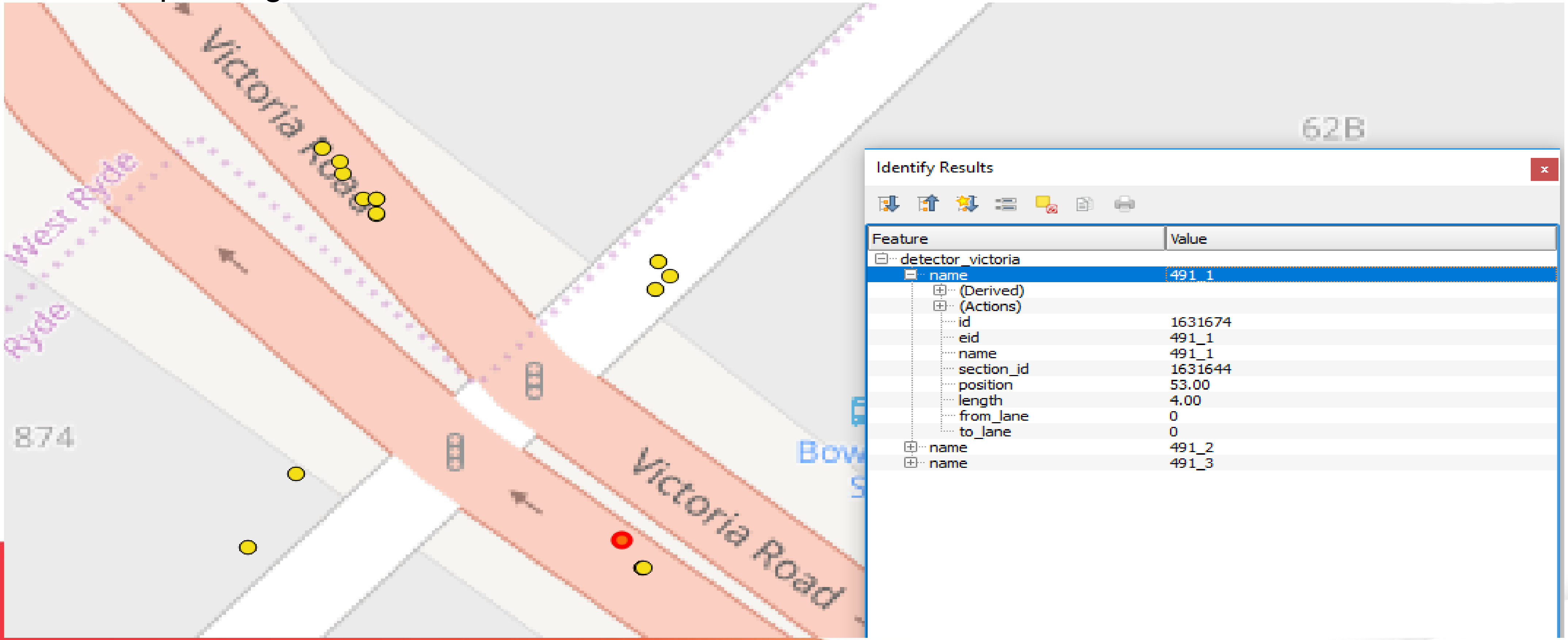
- SCATS traffic counts available for the whole 2017 on all road sections in the Victoria Subnetwork area – 15’ frequency.
- The data from each detector is summed according to their installation location to generate the flow data for the corresponding road section.
- not all sections are equipped with SCATS detectors. There are 2,672 road sections in the model, 85 signalized intersections with the adaptive SCATS control system running, and a total of 4,256 SCATS detectors.
- some incidents are reported in locations with no traffic flow information



Victoria Road – Sydney, AU

3.2 Adding in traffic flow features

- The data from each detector is summed according to their installation location to generate the flow data for the corresponding road section.



Victoria Road – Sydney, AU

3.2 Adding in traffic flow features

- 3 measures of traffic flow were used:
    - a) the reported real-time flow from the 15-min time-interval when the incident was reported (TRF),
    - b) the traffic flow corresponding to 1 hour prior to the accident (TFH) and
    - c) the 15-min to 1 hour traffic flow ratio on each section computed as TFR=TRF/TFH, (a TFR =zero => high congestion as the real-time flow decreases considerably close to the accident start-time as compared to the flow 60 minutes earlier).

Victoria Road – Sydney, AU

3.3 Feature Scenario construction

Baseline Feature Set (BFS): uses all the feature information from Table 1

Feature Set A (FSA):
BFS + flow counts from TRFi, TFHi, TFRi, $i \in \{1, N_s\},)$, where N_s = total number of road sections). This resulted in the addition of almost 700 extra features in the model training.

Feature Set B (FSB): BFS + the traffic flow from only the top 5 closest road sections to the incident location (TRFi, TFHi, TFRi, where $i \in \{1, .. 5\}$).

Feature Set C (FSC): BFS + the aggregated traffic flow from the top 5 closest road sections ($TRF_{top\_5} = \sum_{i=1}^{5} TRF_i$,$TFH_{top\_5} = \sum_{i=1}^{5} TFH_i$, $TFR_{\text{top\_5}} = \sum_{i=1}^{5} TRF_i$).

Feature Set D (FSD): BFS + the traffic flow extracted from all the sections in the vicinity of the reported location of the incidents ($TRF_{\text{dv}} = \sum_{i=1}^{N_r} TRF_i$, $TFH_{\text{dv}} = \sum_{i=1}^{N_r} TFH_i$, $TFR_{\text{dv}} = \sum_{i=1}^{N_r} TRF_i$, where
N_r is the total number of road sections in the selected area and
dv is the distance from the location of the incident to the closest road sections; {100m, 200m, 300m, 500m, 600m}

Summary

4. METHODOLOGY

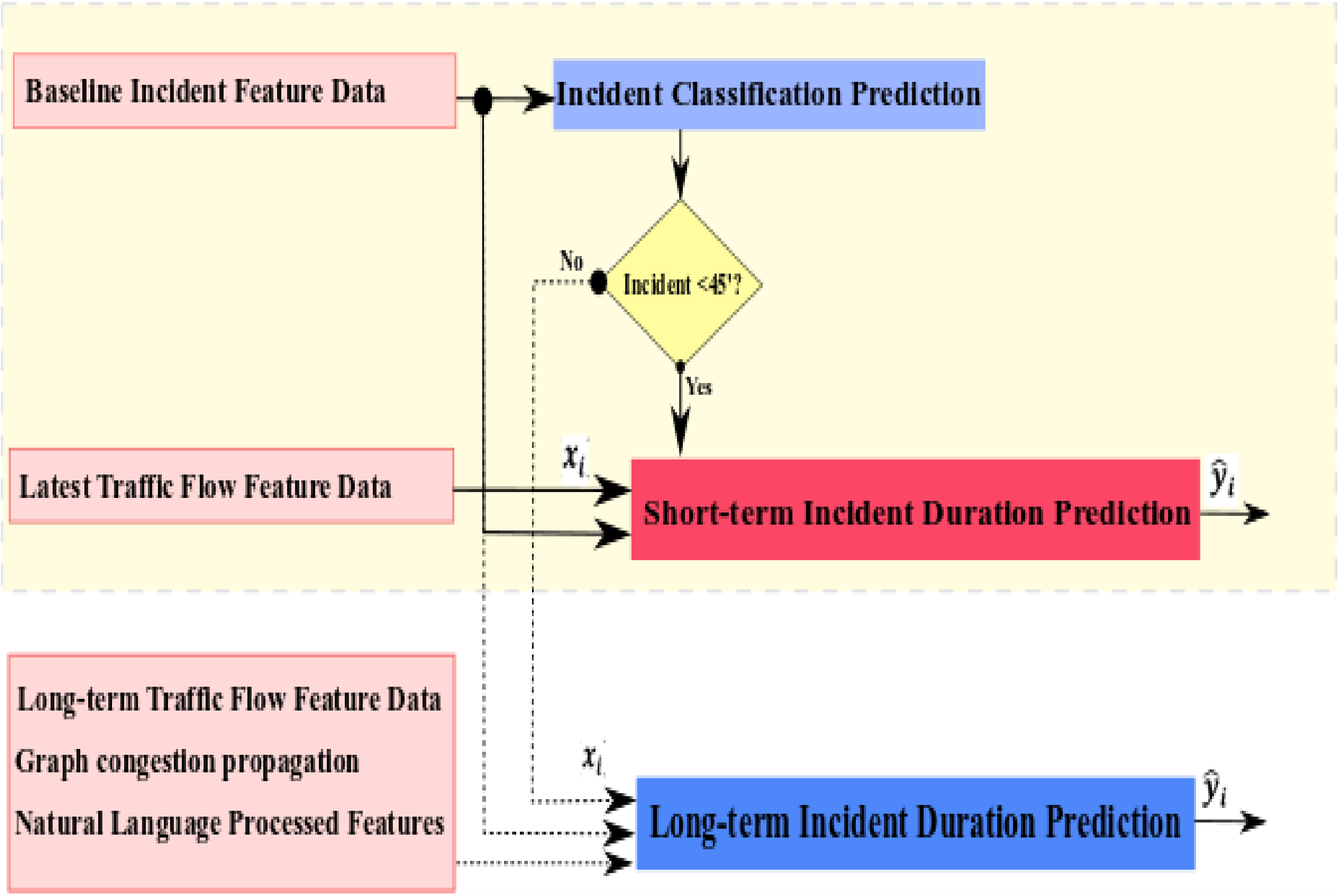


Fig 3 Bi-level incident prediction framework.

4.1 Incident duration classification

$\boldsymbol{X} = [\boldsymbol{x_{i,j}}]_{i=1,..N_i}^{j=1,..N_f}$ the matrix of model features,

where N_i is the total number of incidents used for training the models, and

N_f is the total number of features to be considered.

$\boldsymbol{Y} = [y_i]_{i=1,.N_i}$, Incident duration vector

where y_i is the duration (in minutes) of an incident occurring at a specific time.

The classification problem is to predict $\boldsymbol{Y}$ from $\boldsymbol{X}$, more specifically predict if $\hat{y}_i$ takes one of the following values:

$$\hat{y}_i = \begin{cases} 1, & if\ y_i \le \bar{Y} \\ 0, & if\ y_i > \bar{Y} \end{cases} \qquad (1)$$

where $\bar{Y}$ represents the 45min incident clearance time.

4.1 Incident duration classification

Methods:

- **k-nearest neighbours (kNN)**[14] - doesn’t require specific assumptions about the data distribution or characteristics of the variables to learn,
- **logistic regression (LR)** [13] - focuses on the conditional probability distribution of the predicted variable given its set of features,
- **random forests (RF) [23]** – randomly selects observations/rows and specific features/variables to build multiple **decision trees** and then average the results.
- **gradient-boosted decision trees (GBDT)** [16] – from the entire dataset, uses all the features/variables of interest to build decision trees where the leaves are the final predicted class.
- **extreme-boosted decision trees (XGBoost)** [17] – enhanced version of GBDT with a regularization parameter in the objective function.

[13] G. Valenti, M. Lelli, and D. Cucina, "A comparative study of models for the incident duration prediction," *European Transport Research Review,* vol. 2, no. 2, pp. 103-111, 2010/06/01 2010.
[14] Y. Wen, S. Y. Chen, Q. Y. Xiong, R. B. Han, and S. Y. Chen, "Traffic Incident Duration Prediction Based on K-Nearest Neighbor," *Applied Mechanics and Materials,* vol. 253-255, pp. 1675-1681, 2013.
[16] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 9, pp. 2303-2310, 2017.
[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.
[23] K. a. A.-R. Hamad, Rami and Zeiada, Waleed and Dabous, Saleh Abu and Khalil, Mohamad Ali, "Predicting Incident Duration Using Random Forests," presented at the Transportation Research Board 97th Annual Meeting, Washington D.C., 2018.

4.1 Incident duration classification

We perform a five-fold cross-validation (5CV):

- the dataset is randomly divided into five subsets (or folds), each containing the same proportion of the positive and the negative class (e.g. stratified folds).
- Iteratively, each fold serves as a test set, while the remaining four folds are used as training set.
- The model parameters are fit on the training set, and the predictions of the incident durations are obtained for the test set.

Evaluation:

$\boldsymbol{A} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$: ratio of correct predictions over all prediction; sensitive to class imbalance

$\boldsymbol{P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$: how many of the predictions made by the learner are correct

$\boldsymbol{R} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$: how many of the correct (e.g. true) examples were correctly predicted by the learner

$\boldsymbol{F_1} = \frac{\boldsymbol{2PR}}{\boldsymbol{P+R}}$: mean of Precision and Recall; to maximize F1, the learner needs to have simultaneously a high Precision and a high Recall

4.1 Incident duration regression

XGBoost enhances GBDT (which sequentially processes a combination of trees from weighted training data with a slow learning rate) by:

- introducing a regularization parameter in the learning objective function (to control over-fitting),
- Adapting to parallel tree learning through sparsity-aware capability
- having a better support for multicore processing which reduces computational time

Given a dataset $D = \{(x_{ij}, y_i)\}$, where $|D| = N_i$ and $\{x_{ij} \in \mathbb{R}^{N_f}, y_i \in \mathbb{R}\}$, the XGBoost model uses ***K*** additive functions to predict the incident duration as:

$$\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \emptyset(\boldsymbol{x}_{\boldsymbol{ij}}) = \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}}),\ \boldsymbol{f}_{\boldsymbol{k}} \in \boldsymbol{\mathcal{F}} \qquad (4)$$

where ***K*** is the number of generated trees, and $\boldsymbol{f}_{\boldsymbol{k}}$ are functions in the functional space $\boldsymbol{\mathcal{F}}$ defined as:

$$\boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}}) = \omega_{\boldsymbol{q}(\boldsymbol{x})}, \omega \in \mathbb{R}^{\boldsymbol{T}}, \boldsymbol{q}\colon\ \mathbb{R}^{N_f} \rightarrow \{1,2,..T\},$$

T is the number of leaves in the tree; each $\boldsymbol{f}_{\boldsymbol{k}}$ corresponds to an independent tree structure ***q*** and leaf weights $\omega_{\boldsymbol{q}(\boldsymbol{x})}$. The objective function to be minimized is given by:

$$\boldsymbol{\mathcal{L}}(\emptyset) = \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \boldsymbol{l}(\boldsymbol{y}_{\boldsymbol{i}}, \widehat{\boldsymbol{y}}_{\boldsymbol{i}}) + \sum_{\boldsymbol{k}} \Omega(\boldsymbol{f}_{\boldsymbol{k}}) \qquad (5)$$

where $l(y_i, \hat{y}_i)$ is a MAPE function and Ω is the regularization term which penalizes the complexity of the model and is expressed as:

$$\Omega(f_k) = \lambda T + \frac{\gamma}{2}\|\omega\|^2$$

4.1 Incident duration regression

Performance evaluation:

- $\boldsymbol{MAPE} = \frac{\mathbf{100\%}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left| \frac{\boldsymbol{y_i - \widehat{y_i}}}{\boldsymbol{y_i}} \right|$ (6)
- $\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$

Hyper-parameter tuning through randomized search

- **Randomized-Search**:
  - selects randomly a (small) number of hyper-parameter configurations to use through cross-validation.
  - for a high enough number of random samples (e.g. 100-200) the random search was faster than **grid-search**, **Bayesian optimisation**

- For our cross-validation setup, we train each learner for:
  - 10 times (number of learning folds) x
  - 5 times (number of hyper-parameter tuning folds) x
  - 500 (number of random hyper-parameter combinations) = 25,000 times.
- Total execution time = 10 minutes, on a computational machine with 24 cores.

4.1 Feature Importance classification

Selecting important features:

- the influence of each feature data set can be very different on the prediction results.
- not all features are efficient for improving the result accuracy!
- We need a fast metric which drops unnecessary information (we have scenarios with almost 800 features!).

Shapley value :

- originates in the game theory: where a coalition ***S*** of ***n*** players (belonging to a set ***N)*** need to cooperate to obtain an overall gain, called the worth of the coalition ***v(S)***.Main question: how to divide the gain among players based on contributions?
- Shapely is a fair distribution of gain among the players, assuming they all cooperate. For our prediction *p*, the Shapley value for a specific feature *i* (out of the total N_f features) can be expressed as:

$$\emptyset_i(p) = \sum_{S\subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \boxed{(v(S \cup \{i\}) - v(S))} \quad (8)$$

what would the prediction of the model be without a feature *I?*

- Practically the Shap value calculates the marginal contribution of a feature i to the entire feature set over the number of features excluding i.

Results

Incident Classification:

Accuracy:
- kNN ranks lowest at 65%,
- RF and GBDT : highest score of 69% but RF is overfitting (testing is higher with 21% of the);

F1 :
- GBDT has the highest F1 score in testing (82%)
- GBDT - close to XGBoost

Recall:
- Aside from kNN, all other four ML models have a testing Recall score above 95% - that they are highly capable of identifying incidents <45 min.

Precision :
- RF, GBDT and XGBoost – achieve the highest performance across all methods
- Even though their testing performance are similar, RF appears to over-fit the training data more than the other two models

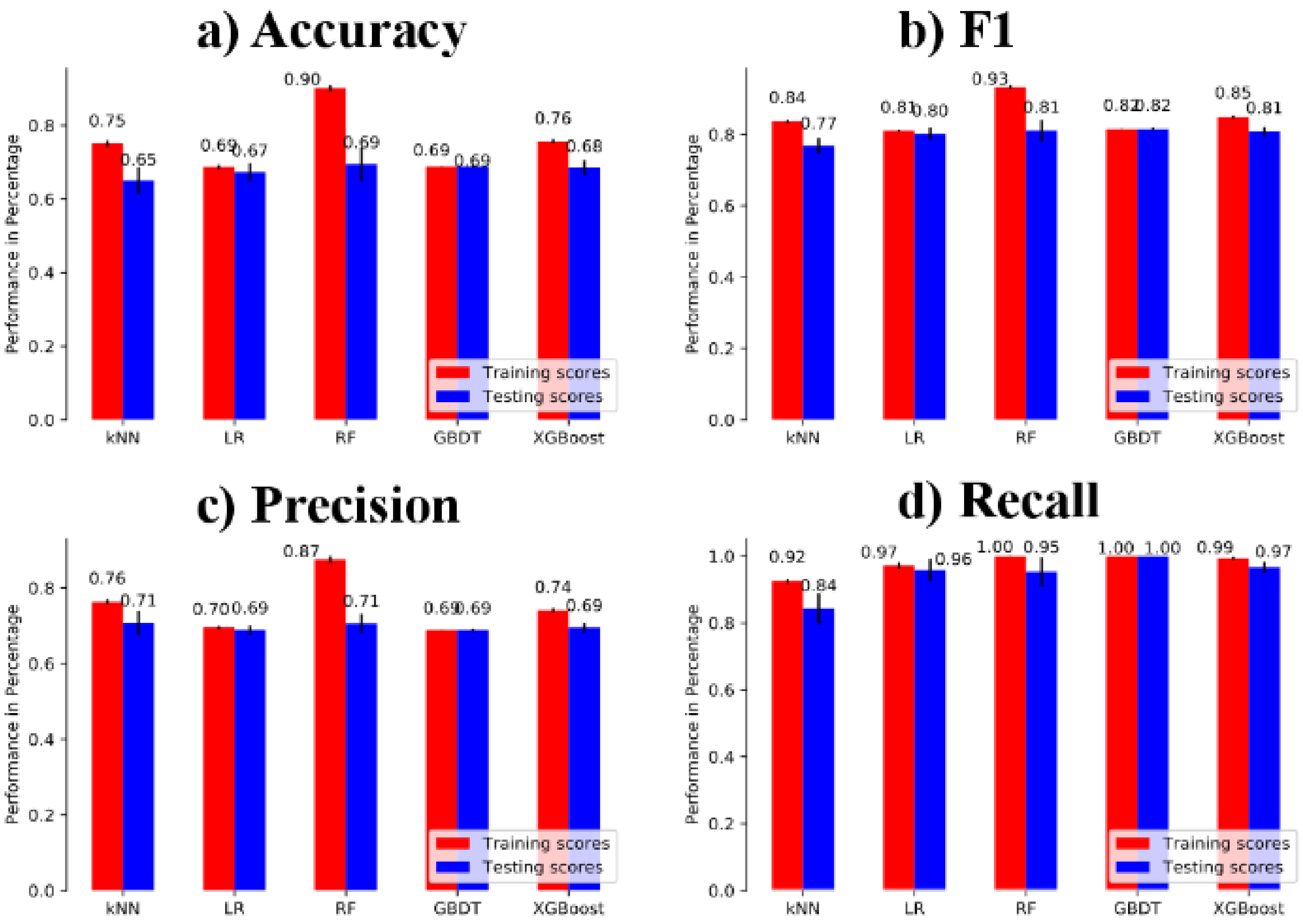


Fig 5 Performance comparison of different classification algorithms on the Baseline Feature Set: a) Accuracy b) F1 c) Precision d) Recall.

Results

Incident Regression:

- **Zero or very small durations** (<1-2min) are noisy for the model training and prediction.

- Removing these outliers reduced the MAPE from 221.04 to 120.34 for GBDT and from 77.84 to 68.77 for XGBoost

- **Large outliers** – tested the log space transform: GBDT this procedure reduces the error considerably (from 120.34 to 83.26 after outlier removal),
- this has limited impact for the extreme boost model most likely due to its additional regularization terms.

OBS: the modelling results presented in the following have a learning in the **original incident duration space** and are **trained without the previously mentioned outliers**.

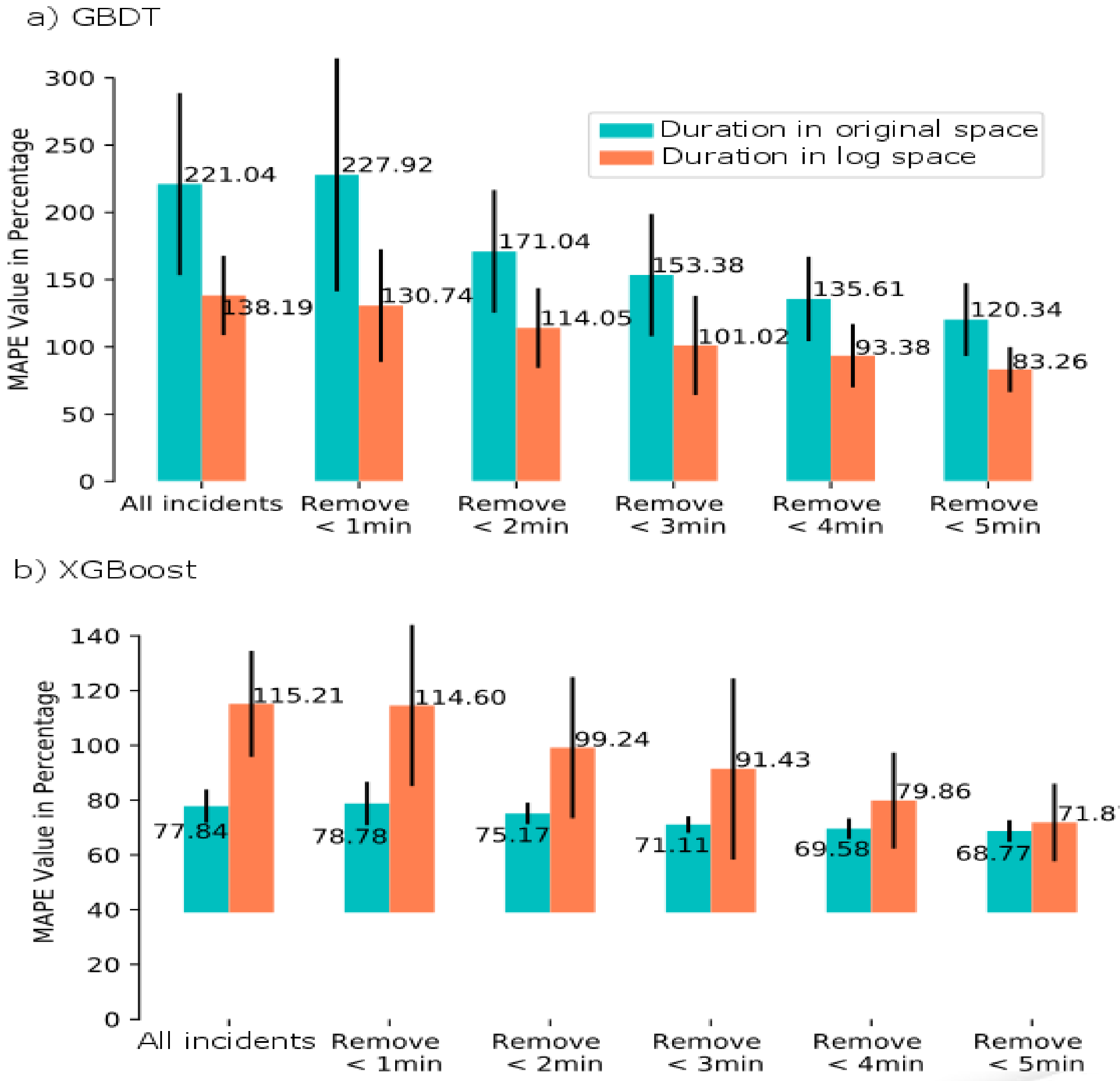


Fig 6 MAPE comparison among datasets with incremental outliers removed.

Results

Incident Regression:

Four regressors (3 training measures)

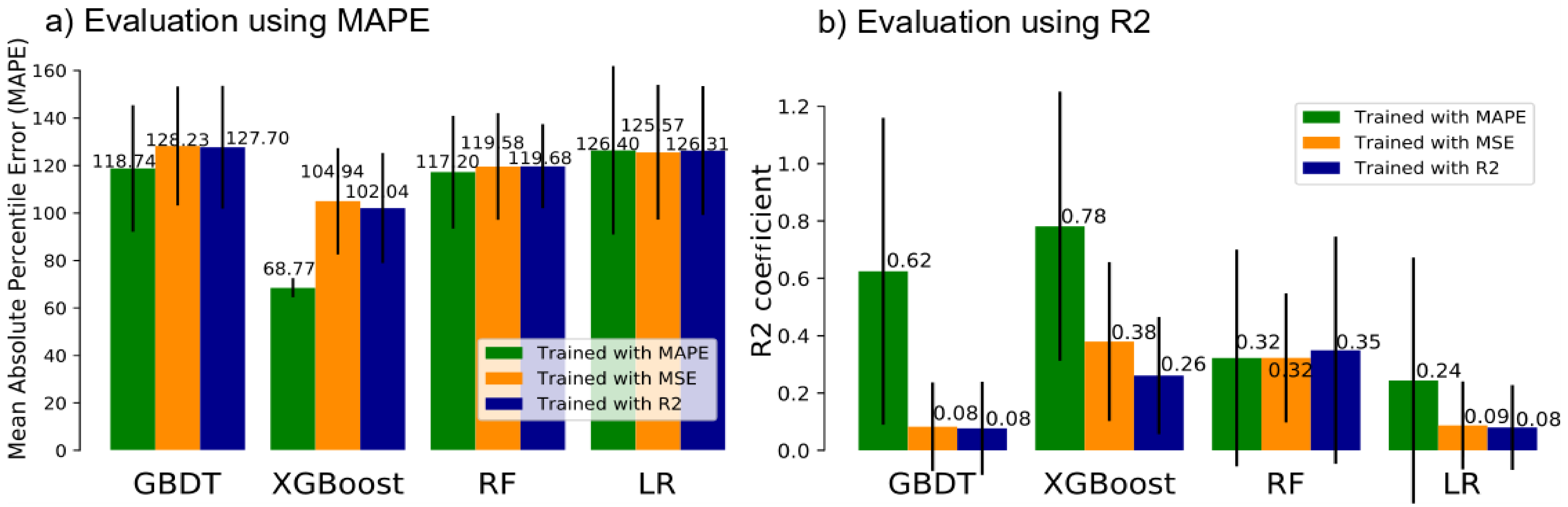
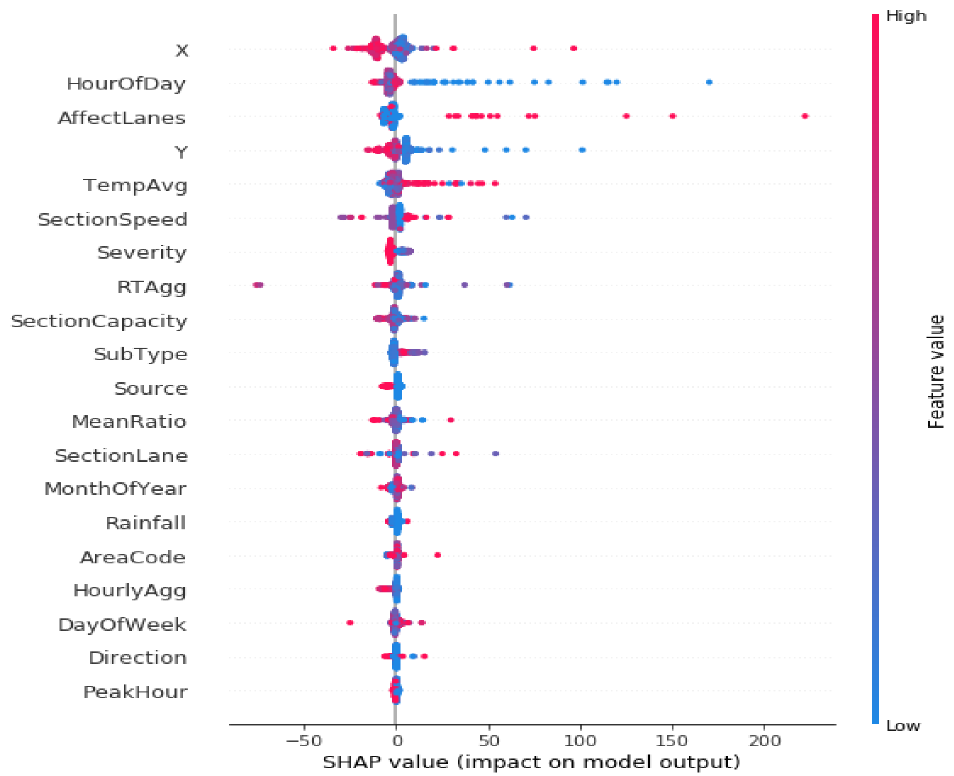
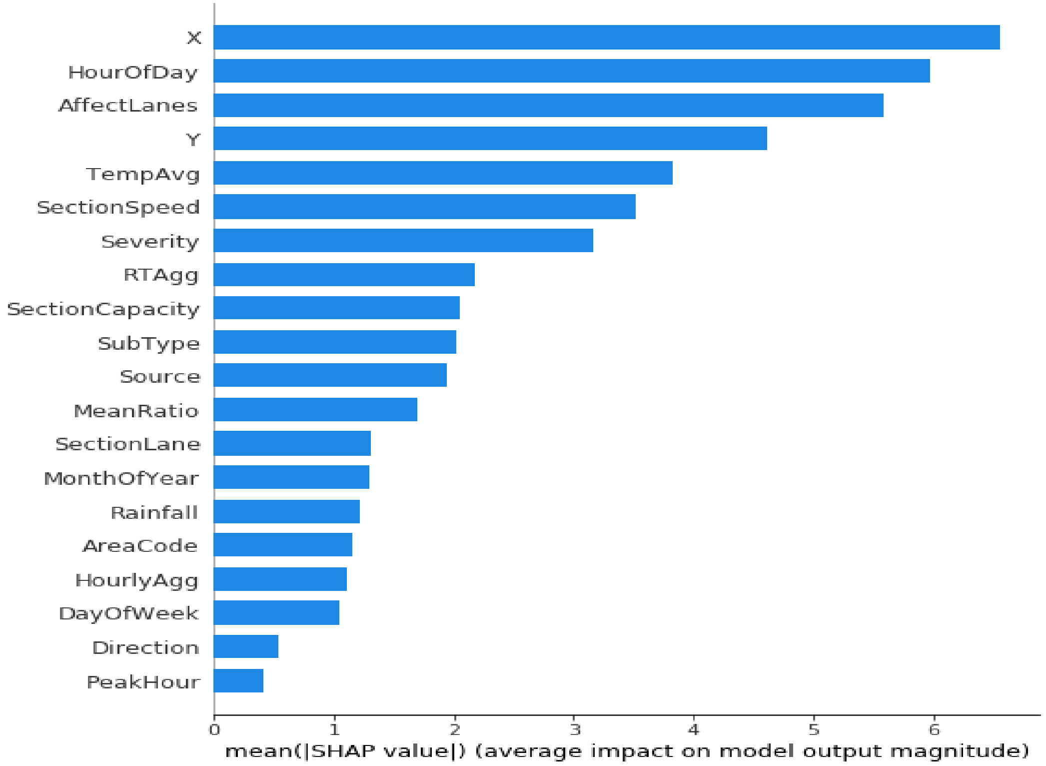


Fig 6 MAPE comparison among datasets with incremental outliers removed.

Results

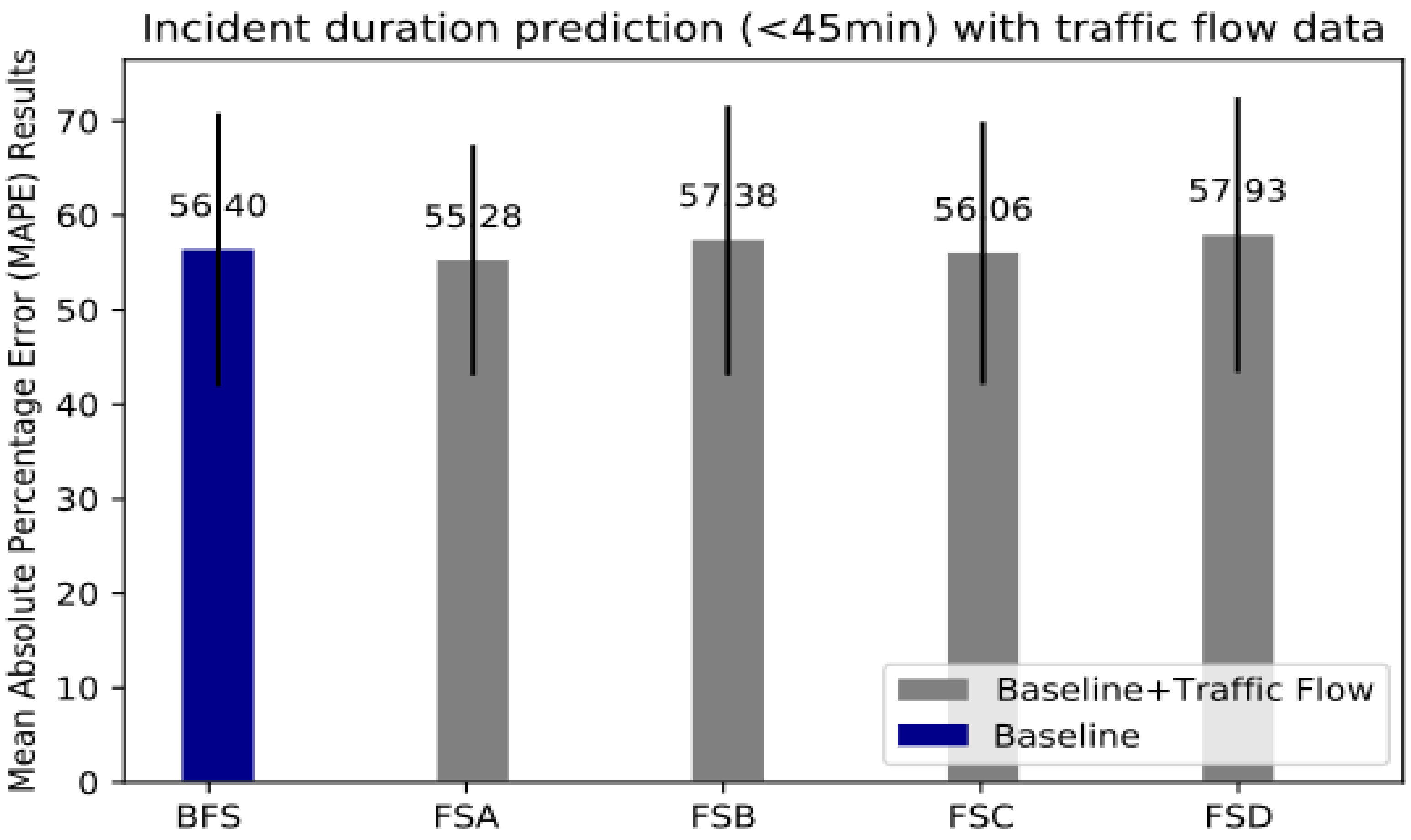
Feature importance





Results

Scenario evaluation for traffic flow impact analysis:



AP 2020
ITS Asia Pacific
Forum **Brisbane**

17th Asia Pacific Forum on Intelligent Transport Systems

Brisbane Convention & Exhibition Centre
Queensland, Australia | 25–28 May 2020

REGISTRATION OPEN NOW

itsaustralia
Intelligent Transport Systems

itsap2020.com

THANK YOU!

adriana-simona.mihaita@uts.edu.au