CRASH Predictive Analytics: A Machine Learning Approach to Improve Highway Safety

Sayyed Mohsen Vazirizade[1], Ayan Mukhopadhyay[1], Geoffrey Pettet[1], Said El Said[2], Hiba Baroud[1], Abhishek Dubey[1]
[1]Vanderbilt University, [2]Tennessee Department of Transportation

Motivation

- Crash injury is among the top ten leading cause of death worldwide. Each year, 1.35 million people are killed from roadway crashes (Global Status Report on Road Safety).
- Risk reduction and effective emergency response strategies can prevent roadway crashes and reduce injury.
- The success of these strategies relies heavily on the ability to anticipate the occurrence of crashes and the determination of risk factors (Fig. 1).

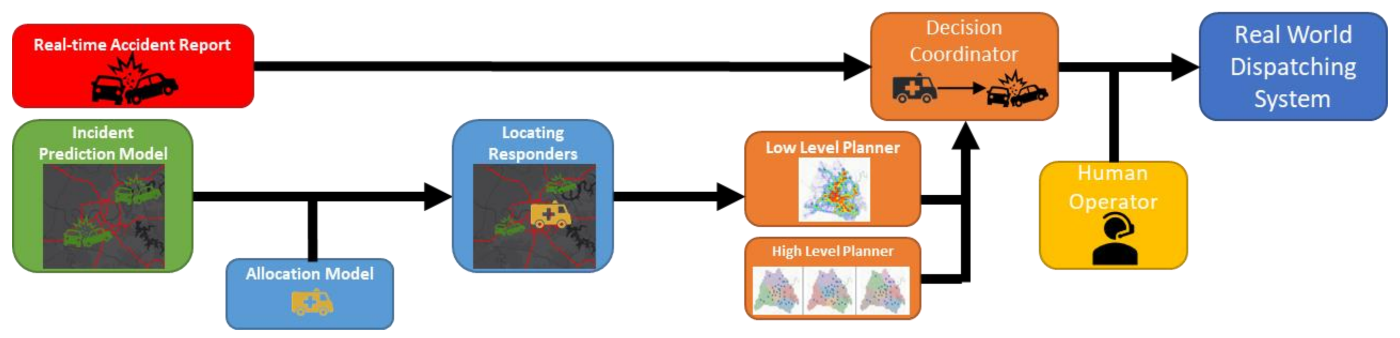
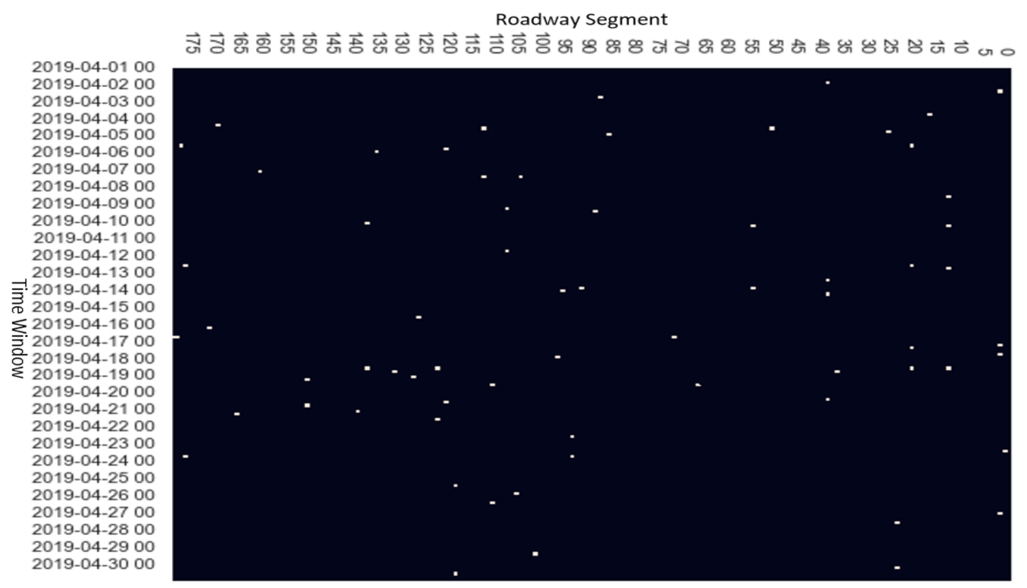


Fig. 1: Typical emergency dispatch model. This research focuses on the prediction.

Objectives

- Design a pipeline to collect and process data from various sources and apply machine learning algorithms to predict the occurrence of accidents.
- Develop metrics to evaluate the performance of machine learning models in improving emergency response.
- Analyze features to determine accident risk factors.



While the frequency of crashes is high, the incidents are rare at large spatial-temporal scales, and the data is sparse.

Fig. 2: Randomly selected road segments for 4-hour time windows in April 2019. Each pixel in the matrix denotes the presence (white) or absence (black) of an accident.

Methods

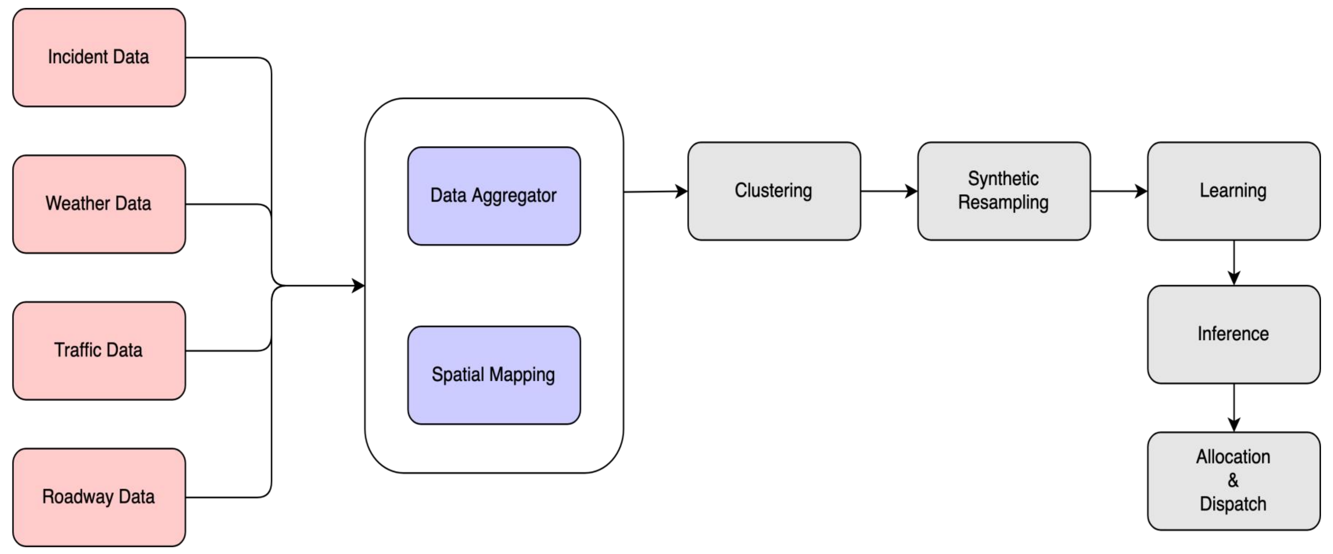
Data

- More than 1.2 TB of data is collected from multiple sources (e.g., INRIX, Weatherbit).
- Features include spatial (no. of lanes), temporal (day of the week), and spatial-temporal (congestion, weather) as well as static and dynamic variables.

Modeling Approach

The goal is to learn the parameters θ of a function, $f\,(X \mid w, \theta)$, over a random variable X (i.e., accident occurrence) conditioned on w (i.e., features).

- *Learning.* Multiple models are considered including Logistic Regression (LR), Neural Networks (NN), Random Forests (RF), and Zero-Inflated Poisson (ZIP).
- *Resampling.* To address the sparsity and imbalance in the data, we perform synthetic random under-sampling (RUS) and random over-sampling (ROS).



Model performance Metrics

Performance is based on emergency response improvement (e.g., response time) and a function of the number of resources (p) and a hyperparameter (α) that controls the penalty on the responders' increased load (Fig. 3).

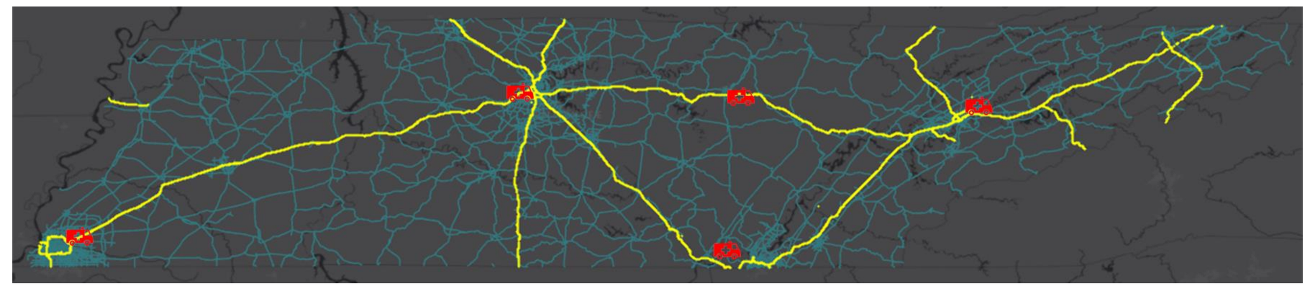


Fig. 3: TN's roadway network (blue), interstate highway (yellow), and potential locations of responders (red vehicles).

Results

Clustering

The data is clustered according to different variables and using different methods (K-means, agglomerative clustering) and a model is developed for each cluster.

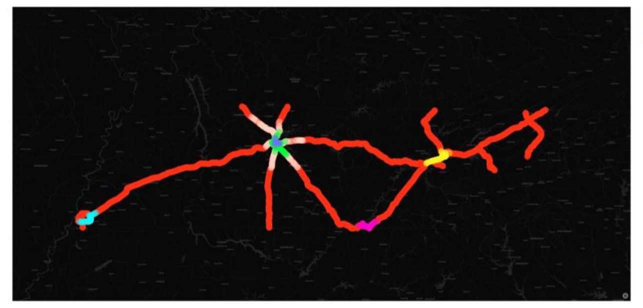


Fig. 4: Road segments are clustered according to the rate of accidents and their connectivity and proximity. This figure is showing 6 clusters.

Model performance

The table below summarizes the different performance metrics used to evaluate the models.

						Average no. of unattended accidents					
						p=10			p=15		
Model	**Resampling**	**Acc.**	**Prec.**	**Rec.**	**F1**	α=0	α=1	α=2	α=0	α=1	α=2
Naïve		95.5	3.8	4.2	4.0	0.59	0.51	0.53	0.06	0.05	0.05
LR	No resampl.	93.0	12.5	30.9	17.7	0.56	0.48	0.51	0.06	0.05	0.05
	RUS	92.3	12.1	34.4	17.8	0.60	0.49	0.52	0.06	0.05	0.05
	ROS	92.4	12.2	34.2	17.9	0.60	0.50	0.52	0.06	0.05	0.05
NN	No resampl.	95.0	19.0	31.6	23.7	0.50	0.48	0.51	0.04	0.04	0.04
	RUS	94.7	18.4	32.7	23.3	0.51	0.47	0.50	0.05	0.04	0.05
	ROS	94.7	18.3	33.1	23.3	0.51	0.47	0.50	0.05	0.05	0.04
RF	No resampl.	95.1	18.9	30.5	23.2	0.58	0.48	0.51	0.05	0.05	0.04
	RUS	95.0	19.4	32.5	24.2	0.53	0.47	0.50	0.05	0.04	0.04
	ROS	95.1	18.3	28.7	22.2	0.58	0.48	0.50	0.05	0.05	0.04
ZIP	No resampl.	93.1	13.1	31.9	18.5	0.54	0.47	0.49	0.06	0.05	0.05
	RUS	93.0	12.7	30.8	17.8	0.62	0.51	0.51	0.05	0.05	0.05
	ROS	93.0	12.8	30.9	18.0	0.63	0.52	0.51	0.05	0.05	0.05

Feature analysis

Visibility is among the most important features and risk factors for highway accidents.

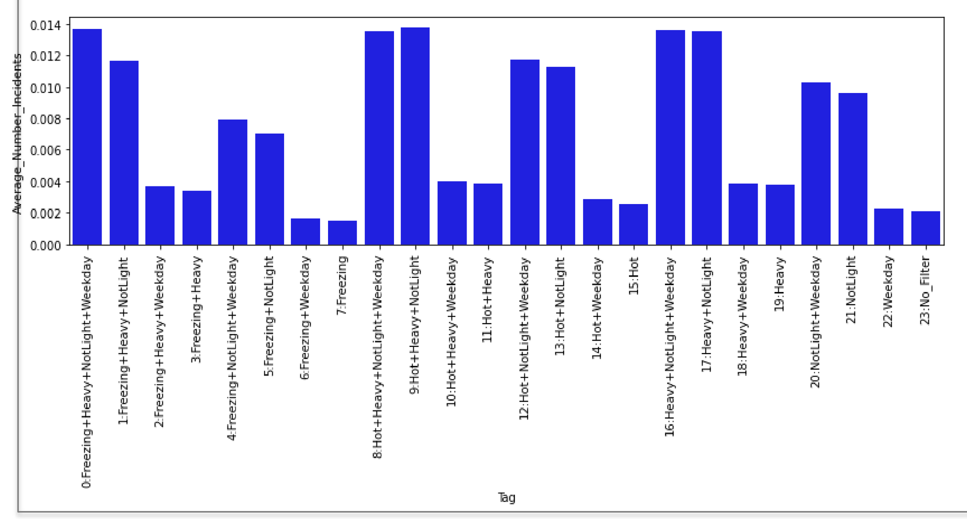


Fig. 5: Feature analysis reveals that specific combinations of feature categories influence the occurrence of accidents. For example, heavy precipitation combined with congestion is a significant risk factor.

Conclusions

- Conventional accuracy metrics can be misleading when data is sparse and imbalanced.
- Improving the predictive accuracy of accident forecasting models improves the emergency response.

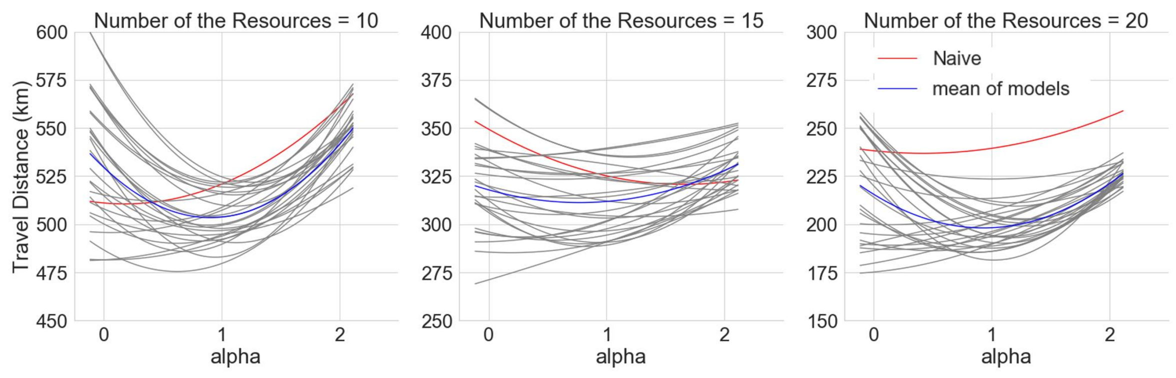


Fig. 6: Performance of models based on distance traveled by emergency responders under different assumptions of p and α.

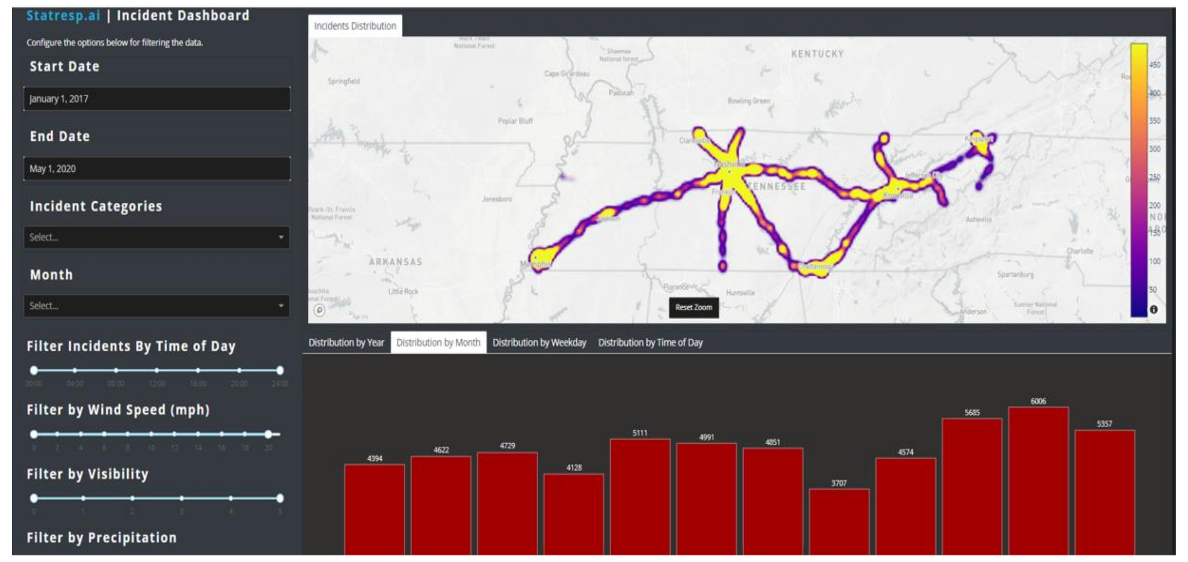


Fig. 7: Dashboard to improve the user experience of the pipeline and support emergency response decisions.

References & Acknowledgements

This work is supported by the Tennessee Department of Transportation and Vanderbilt University.

[1] Vazirizade, S.M., Mukhopadhyay, A., Pettet, G., Said, S.E., Baroud, H. and Dubey, A., 2021. Learning Incident Prediction Models Over Large Geographical Areas for Emergency Response Systems. arXiv preprint arXiv:2106.08307.
[2] Mukhopadhyay, A., Pettet, G., Vazirizade, S.M., Lu, D., Jaimes, A., El Said, S., Baroud, H., Vorobeychik, Y., Kochenderfer, M. and Dubey, A., 2022. A review of incident prediction, resource allocation, and dispatch models for emergency management. Accident Analysis & Prevention, 165, p.106501.

For more information, please visit:
tn.statresp.ai
Nashville.statresp.ai

Contact information:
hiba.baroud@vanderbilt.edu
abhishek.dubey@vanderbilt.edu