Collaborative Sciences Center for ROAD SAFETY

RESEARCH BRIEF

Machine Learning Tools for Informing Transportation Technology and Policy

Little work has been conducted to study practitioner-induced subjectivities introduced into machine learning applications, which is important in understanding causes, influences, and methods for avoidance, particularly in transportation settings. To help fill this gap, this study uses two transportation datasets to examine car and pedestrian crash fatalities, deploying two different machine learning techniques of low and high complexity (logistic regression and neural networks), shown in terms of their relative complexity below.

Complexity map of common machine learning models.

Using both large and small transportation data sets with various features to predict driver and pedestrian injuries and fatalities, the results of these models are compared with one another, as well as with different possible interpretations of the neural network features. The results demonstrate that both the type of model and feature interpretation method produce different results in terms of model performance and assessment of feature importance.

These modeling results highlight several points of human-inserted subjectivity, which include:

- Picking the model to be used,
- Picking stopping rules for training performance factors (software dependent),
- Selecting data training and testing ratios,
- Picking threshold values between fatal and non-fatal values,
- Choosing thresholds between important and unimportant features, and
- Deciding across models with slightly different results, what the actual important features were.

These modeling results highlight other core issues not often discussed in practical applications of these methods, including issues with imbalanced data, which occurs when one class of data (e.g., non-fatalities in the data set utilized) significantly dominates over another (e.g., fatalities).

Another significant issue with the use of such powerful but potentially brittle analytical tools is a lack of checks and balances for results generation and interpretation. There is a very real possibility that decisions could be made from results generated by models that are not exactly wrong but also are not completely correct. Such inherent data-analytic weaknesses need to be accounted for when policymakers make decisions based on machine learning-generated results.

PRINCIPAL INVESTIGATOR

Mary Cummings
DUKE UNIVERSITY

LEARN MORE
www.roadsafety.unc.edu/research/projects

THE UNIVERSITY OF TENNESSEE KNOXVILLE · Berkeley UNIVERSITY OF CALIFORNIA · FAU FLORIDA ATLANTIC UNIVERSITY · Duke UNIVERSITY · THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

This project was supported by the Collaborative Sciences Center for Road Safety, www.roadsafety.unc.edu, a U.S. Department of Transportation National University Transportation Center promoting safety.