An ensemble learning method for variable selection and its application on Railroad Fatal Accidents.

Jing Chen1*, Yan Yan1, and Xi Chen2

1Department of Mathematics and Physics, Wuhan Institute of Technology, Wuhan, China
2China Gezhouba Group Eco-Environmental Engineering Company, Wuhan, China
*Corresponding author’s e-mail: 19090801@wit.edu.cn

Abstract. Scalars based on linear and generalized linear models are commonly used in fatality and disease predictions. Currently, standard approaches for variable selection in linear and generalized linear models are not explored well with high dimensional data. This manuscript proposed a new variable selection method for logistics models with inspiration from heterogeneous ensemble methods. The goal of this study is to extend classical variables selection methods such as stepwise, lasso, or RF in the case of high dimensional accident data with missing values. Real world accidents-collated data sets are collected from the official website of US Railroad Administration and US Police Department from 2000 to 2015. Preprocessing includes log transformation and z score transformation. Simulated sampling was applied due to the imbalanced outcome. The base variables was selected from the Empirical Model (EM) [1], Lasso model[2], Xgboost[3] and the traditional stepwise backward. An intermediate selection was proposed by ensemble training of base variables. Final models built by variables selected improved Area Under the Curve (AUC) from 0.79 to 0.84 and from 0.79 to 0.84 in test data compared to arbitrary APS model. Comparison with the existing model shows that the logistics model based on selected variables has high accuracy.

1. Introduction
In recent years, the proliferation of data makes high-dimensional data popular, which brings new directions and challenges to public health and social research and traditional statistical methods. Accidents happened in digital age, which data can be collected unprecedented instantly, has become a valuable resource. Therefore, how to use the current data in a timely manner is an important support to accident prevention and control.

In railroad related fatalities, grade crossing accidents represent a significant proportion. In 2012 a total of 1,967 at-grade railroad crossing accidents were reported by US Federal Railroad Administration, which lead to 233 fatalities and 941 injuries. Several reports from US Department of Transportation showed most possible cause are drivers’ error such as ignorance and distraction[4]. Hence, common mitigation strategies can be categorized into three categories: engineering, education and enforcement helps on driver’s decision-making [5,6]. To mitigate grade crossing accidents, a taxonomy is necessary to categorize accidents: unsafe acts, individual differences, train visibility, passive signs and active warning systems, and physical constraints [7]. Little information exists describe the vehicle drivers involved in the accidents, and their reaction to the current mitigation strategies.

FRA presented several prediction models in the past, and the most recent one is in Dec, 2020 [1].
However, all the information used there are from FRA datasets. To fully understand drivers involved and their feedbacks to current mitigation strategies, multiple data sources are in need to provide a robust accident mechanism and confirmation system. But when including more possible datasets, the dimension of the data also increases. What is more, scalar score is a popular risk and disease evaluation method that ease the work of frontline responders. Most of these scalar score system are built from linear prediction models. However, current variable selection methods for linear models are not explored well.

In the past research, the commonly used variable selection methods include subset selection, shrinkage methods and nonlinear tree based machine learning algorithm [8]. The emerging ensemble learning algorithms contributes to the stability of variable selection in high dimensional data, ans has been proved to be successful in various applications [9-11]. The manuscript presents a novel variable selection design from high dimension data, to eventually build a logistics prediction model. The innovation of this paper is: (1) take vehicle driver’s risk behavior and social status variables into account; (2) explore variable selection methods for logistics prediction models by ensembling approach (EL).

2.Material and Method

2.1. Experiment Data

As in Fig1, the whole dataset in this research come from three public datasets: U.S. Census Bureau (CENSUS), Fatality Analysis Reporting System (FARS) and Federal Railroad administration (FRA). CENSUS provided some general population characteristic variables including mean education level, mean population density, mean household income and so on. FARS provided fatality related data including the outcome variable has fatal accidents or not. FRA provided railroad crossing information as well as railroad crash report, with all information of the crash, train, vehicle, and risk factors of all 419,164 grade crossings.

![Figure1. Data Collection Process Flowchart describing the data collection process.](image)

2.2. Ensemble method

Fig 2 showed an overview of the ensemble integration approach. First, data was randomly separated into train and test datasets with a 7:3 ratio. Due to the relatively rare fatal cases, a bootstrapping based resampling method was applied to balance train dataset [13]. Then, local models were trained each with one of the three standard variable selection algorithms, including LASSO, stepwise method (STEP), and Xgboost (XGB). Next, to find the optimal combination of selected variables, the super learner algorithm was introduced to minimizing the cross-validated risk [SL]. The whole procedure was conducted with 10-fold nested cross validation to reduce overfitting and improve evaluation result.
Figure 2: Overview of the Ensemble approach for logistics models (EL). The full dataset is integrated and resampled to several balanced simulated datasets. One base predictor each is trained on the individual data sets using three variable selection algorithms including LASSO, stepwise, and XGBoost with variable importance. The variables selected by these base methods are then combined with variables from arbitrary prediction model [1], applied to logistics models and ensembled by super learner algorithm.

2.3. Evaluation Methodology
The performance of the variable selection result was evaluated by the result of predicting test dataset with logistics models. Each logistics model was built with different variable selection result. We compared the representative model in the following three categories: the sensitivity, specificity, and mean squared error of the prediction.

3. Result
Table 1 showed risk estimated by mean squared error of prediction result by cross validation of super learner, where LM_MIX is the combination of variables from arbitrary models and ensemble learning results. LM_mixed archived the lowest risk 0.19 followed by three variables selection methods: stepwise/lasso/XGBoost. LM_MIX also presented best AUROC performance of 0.84. SAR is the score combining performance [14], and LM_MIX also shows best result. It is notable that all logistics models built from variable selected by ensemble learning performs better that arbitrary model, which reaffirming the utility of these ensembles for predictive logistics models.

| Model  | Risk | B.Acc | Acc | Sen | Spe | MSE | FS | Roc | SAR |
|--------|------|-------|-----|-----|-----|-----|----|-----|-----|
| LM_MIX | 0.19 | 0.75  | 0.84| 0.86| 0.63| 2.07| 0.91| 0.84| 1.27|
| LM_STEP| 0.2  | 0.75  | 0.84| 0.86| 0.64| 1.49| 0.91| 0.82| 1.07|
| LM_lasso| 0.2  | 0.73  | 0.77| 0.78| 0.67| 0.96| 0.85| 0.8   | 0.87 |
| LM_XGB | 0.2  | 0.73  | 0.76| 0.77| 0.69| 1.14| 0.84| 0.8   | 0.93 |
| LM_AM  | 0.21 | 0.72  | 0.77| 0.79| 0.65| 0.75| 0.86| 0.79 | 0.80 |

Also, the variables from empirical model LM_AM are all selected by basic selection algorithms, and maintained well in the ensemble integration process. The LM_MIX performs better than traditional LM_AM. This improvement is likely due to introducing more vehicle driver side variables in the model. It was reported that driver’s error can relate with more than half of the grade crossing accident and the variable selection result back up the theory. Also, local population density showed significant relationship with the fatality.

The odds ratio of core variables across logistics models were in Figure 3. As it showed, Max Speed Limit and Day Through Trains are variables maintain in all models, and also related to an increase risk of fatal accidents. Driver’s Education Level only appeared in LM_MIX model, which might related to the increase of predication accuracy. From this figure, the highway pavement and gates could relate to
a decreased risk.

![Figure 3: Odds ratio of core variables across four logistics models: LM_STEP, LM_Lasso, LM_XGBoost, LM_MIX.](image)

4. Discussion
High dimensional data is one of the main challenges for public health researches at the digital era, especially when a final logistics model is in need. In this manuscript, we proposed a novel approach, named ensemble for logistics (EL), for variable selection in the framework combined common variable selection methods and logistics models. Through the problem of grade crossing fatality prediction, we demonstrated that EL performs better than common variables selection methods and Empirical formula.

This proof of principle study has limitations that offer possible direction for future works. First, also the algorithm can be easily adapted in the case of mixed models, additional work has to be done to tune the parameters. We could know the whole story better if we introduce hierarchical model structure. Also, we did not explore the specific case of data missing not at random. In this case, the algorithm need to be adapted by suitable imputation method. Moreover, vehicle-driver related variables were from police report, which may introduce some bias. A future state-level mixed model might be helpful to improve accuracy.

Acknowledgement

1Scientific Research Foundation, Wuhan Institute of Technology. (000017-20QD48)

Reference

[1] Daniel Brod and David Gillen. (2020). A New Model for Highway-Rail Grade Crossing Accident Prediction and Severity. United States Department of Transportation, Federal Railroad Administration, OMB No. 0704-0188, MH-2004-065

[2] Robert Tibshirani (1994). Regression Shrinkage and Selection Via the Lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 58, 267–288.

[3] Tianqi Chen and Carlos Guestrin (2016). XGBoost A Scalable Tree Boosting System. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August 2016, 785–794.

[4] Office of the Inspector General (2004), United States Department of Transportation, Audit of the Highway-Rail Grade Crossing Safety Program, MH-2004-065, June 16, 2004, p. 2.

[5] Sussman, Donald & Raslear, Thomas. (2007). Railroad Human Factors. Reviews of Human Factors and Ergonomics. 3. 148-189. 10.1518/155723408X299870.
[6] Jeng, O.J.. (2005). Survey of Driver Perceptions of Railroad and Light Rail Warning Devices/Grade Crossings. FHWA-NJ-2004-025.

[7] Caird, J. K. , Edwards, C. J. , Creaser, J. I. , & Horrey, W. J. (2002). Contributing factors to accidents at intersections for older drivers (Report No. TP13939E). Montreal, Canada: Transport Development Centre, Transport Canada.

[8] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer.

[9] Mohapatra S, Patra D, Satpathy S (2014) An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. Neural Comput Appl 24:1887–1904

[10] Alaraj M, Abbod MF (2016) Classifiers consensus system approach for credit scoring. Knowl-Based Syst 104:89–105

[11] Rojas-Thomas JC, Mora M, Santos M (2017) Neural networks ensemble for automatic DNA microarray spot classification. Neural Comput Appl. https://doi.org/10.1007/s00521-017-3190-6

[12] van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6, Article25.

[13] Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. The R Journal Vol. 6/1: 79-89.

[14] Caruana, Rich & Niculescu-Mizil, Alexandru. (2004). An Empirical Evaluation of Supervised Learning for ROC Area.. 1-8.