Crash prediction for a french highway network with an XAI-informed Bayesian hierarchical model

Thomas Veran
Data New Road & LIRIS
UMR 5205 CNRS, INSA Lyon, France
thomas.veran@insa-lyon.fr

Pierre-Edouard Portier
LIRIS
UMR 5205 CNRS, INSA Lyon, France
pierre-edouard.portier@insa-lyon.fr

François Fouquet
Data New Road
Lyon, France
ffouquet@data-newroad.com

***Abstract*—Worldwide, highway accidents have important social and financial impacts. Crash Predictions Models (CPM) are used to reduce their frequency and gravity. They belong to two main categories: generalized linear models (GLM) and nonparametric machine learning (ML) algorithms. Broadly speaking, the former offer better interpretability but tend to have worse predictive performances than the latter. However, for highway infrastructures managers, efficient predictions of accident count must come with explanations so as to give rise to efficient safety actions. Therefore, to balance predictive power and interpretability, we propose a methodology that combines Bayesian learning of hierarchical GLM with automatic detection of latent structures and interactions through methods borrowed from the field of explainable artificial intelligence (XAI). Promising results are obtained with experiments conducted on crash count data from 2008 to 2017 on a large part of the french highway network. Moreover, we tested our approach on three public datasets covering a broad variety of contexts in terms of volume, data types and tasks (viz. classification and regression). These experiments confirm that our framework outperforms traditional GLM models while getting close to the best ML models and remaining interpretable.**

***Index Terms*—Machine Learning, model interpretability, SHAP, Bayesian hierarchical modeling**

I. Introduction

According to the World Health Organization [1], approximately 1.35 million people are killed each year on roadways around the world. Related expenses average 3% of the gross domestic product (GDP) of a country. As stated by the French Road Safety Observatory [2], these costs grow exponentially with the severity of the accidents. In 2018, a Property Damage Only (PDO) accident incurred expenses up to $5\,154$ euros while the average cost of an accident with at least one fatality was $3\,360\,000$ euros.

Crash prediction models (CPM) are trained on historical data to estimate the likelihood of future crashes given the values of explanatory variables (e.g., traffic, speed limit, altitude, etc.). They are mainly used to identify risk factors in order to steer the evolution of safety policies. In their survey [3], Lord and Mannering provide a broad perspective on the variety of data-related issues raised by crash count prediction: over-dispersion of count data, temporal and spatial correlations due to multiple measurements of a same location at different times, fixed parameters that cannot adapt from one roadway to the next, low sample-mean due to the sparsity of crashes, non-linear relationships between crash-frequencies and explanatory variables, etc. Most of theses issues are made more prominent with the use of parametric models, in particular generalized linear models (GLM). Indeed, GLMs must undergo many transformations to adapt to the crash count prediction context (e.g., the choice of a non-normal likelihood distribution, the integration of random effects and hierarchical models, etc). Whereas non-parametric approaches (e.g. neural networks, tree-based algorithms, SVM...) will deal with most of these issues without the need for specific adaptations and will usually offer better predictive performances than parametric models. However, Bayesian inference of GLMs is still highly desirable for CPM as it offers many opportunities to understand the relationships between crash frequencies and explanatory variables while non-parametric approaches behave as black-boxes.

Therefore, we propose a data-driven approach to automatically associate a hierarchical structure and a non-linear functional form to a Bayesian inferred GLM for crash count prediction. In the first place, a well-chosen hierarchical structure can handle correlations among groups of observations and significantly improve the quality of the predictions and of their interpretation. However, in the literature, this structure, which is dependent on the dataset, is the result of expert knowledge. On the contrary, we propose to learn it from the data by analyzing the results of a black-box machine learning model with SHAP [4], a game-theoretic approach from the field of explainable artificial intelligence (XAI). Secondly, a GLM regression model is usually obtained by attaching a linear model to a parameter of the likelihood distribution, however a more complex non-linear functional form would often be necessary to take into account interactions between the explanatory variables. Therefore, we again propose to learn the relevant interactions from the data by analyzing the results of a special kind of polynomial neural network with an automatically inferred layered-structure. In this way, we combine the strengths of nonparametric ML algorithms and Bayesian inferred GLM into an efficient and interpretable framework.

We obtain promising results on a crash-count dataset of the AREA french highway network made of more than 430 km of roads in the Rhône-Alpes region and registering $12\,554$ accidents from 2008 to 2017. We also test our approach on three open datasets of varied volumes and covering regression

and classification tasks. In all cases, we outperform classic GLM and we get close to the performance of the best ML models while remaining interpretable.

The article is organized as follows. We start with an overview of the related works for, first, Bayesian inferred GLMs and ML algorithms applied to crash-count prediction and, second, the model interpretability methods, in particular SHAP. Then, we describe our framework in the methodology section. Finally, in addition to providing detailed analysis on the highway network dataset, we share supplementary results for three public datasets before concluding.

II. Related works

In this section, we give an overview of the related works on the application of GLM and ML algorithms to the crash-count prediction problem. We also introduce model interpretability methods that can alleviate the lack of transparency due to the black-box behavior of ML algorithms.

A. Bayesian inferred GLMs

Crash-count data observations y_i are often modeled with a Poisson distribution [5]–[7] (i.e. a special shape of the binomial distribution with a small probability of an event and a large but unknown number of trials). A regression model is obtained by attaching a linear function of the $1 \dots j$ explanatory variables x_{ij} (e.g. traffic, speed limit, etc.) to the rate λ_i of a Poisson likelihood. A log-link function is also necessary to map the positive Poisson rate to the real line covered by the linear model:

$$\begin{aligned} y_i &\sim Poisson(\lambda_i) \\ log(\lambda_i) &= \beta_0 + \sum_j \beta_j x_{ij} \end{aligned}$$

Thanks to the advances of probabilistic programming [8], automatic and efficient Bayesian inference (BI) can be used to compute estimates of the distributions of the parameters given observed crash-count data [9]–[11]. In that case, given a great initial uncertainty, parameters β_j can be assigned zero-centered normal flat priors with a conservative (i.e. large enough) standard deviation. With the alternative frequentist approach, regression coefficients β_j are estimated by maximizing the non-linear likelihood with an iterative algorithm such as the Newton-Raphson method. However, this approach only offers point estimates of the parameters and doesn't permit the introduction of prior information.

For a counting process described by a Poisson random variable, after conditioning on the predictors, the variance equals the expected value. Often, for crash-count data, the observed variance exceeds this amount and a negative binomial (NB) likelihood is used instead of Poisson. Given the probability of a crash, the NB gives the probability of observing n crashes before the α-th non-crash. With λ representing the mean, its probability mass function can be parameterized as follows [12]:

$$\text{NB}(n; \lambda, \alpha) = \binom{n+\alpha-1}{n} \left(\frac{\lambda}{\lambda+\alpha}\right)^n \left(\frac{\alpha}{\lambda+\alpha}\right)^\alpha$$

The variance of the NB is $\lambda + \frac{\lambda^2}{\alpha}$. Therefore, α controls the amount of over-dispersion. When $\alpha \longrightarrow \infty$ the NB likelihood approaches a $Poisson(\lambda)$ distribution. Moreover, it can be shown (see for example [12]) that a $NB(\lambda, \alpha)$ corresponds to a $Poisson(\lambda)$ where λ comes from a $Gamma(a = \frac{\alpha}{\lambda}, b = \alpha)$, with the gamma distribution parameterized as follows:

$$Gamma(x; a, b) = \frac{a(ax)^{b-1}e^{-ax}}{\Gamma(b)} \text{ for } x > 0$$

This Poisson-gamma model – a continuous mixture of Poisson distributions with rates distributed as a gamma distribution – has been identified as a reference model by road-safety experts (see for example the AASHTO Highway Safety Manual [13]). We find it at the core of many related works ([14]–[17]) where crash count regression is done with a linear model attached to the λ parameter of the NB through a log-link function.

GLM can also be refined into multilevel models to take into account clusters of related observations. For example, crashes occurring in a given geographical region may possess specific characteristics while not differing entirely from crashes in other regions. A simple Poisson regression, by pooling all the observations together, would assume an invariant population and couldn't benefit from regional peculiarities. Otherwise, k clusters could be modeled with the addition of $k-1$ mutually exclusive binary variables to the linear model, but this would correspond to no pooling at all and the clusters would be assumed independent of one another. Contrariwise, multilevel models offer partial pooling through an adaptive regularizing prior. Thus, in the following multilevel Poisson regression with k clusters, hyperpriors μ and σ will allow an adaptive shrinkage of the cluster-specific β_{jk} towards a common mean[a].

$$\begin{aligned} y_{ik} &\sim Poisson(\lambda_{ik}) \\ log(\lambda_{ik}) &= \beta_{0k} + \sum_j \beta_{jk} x_{ijk} \\ \beta_{jk} &\sim Normal(\mu, \sigma) \\ \mu &\sim Normal(0, 100) \\ \sigma &\sim HalfNormal(100) \end{aligned}$$

In this way, Jones and Jørgensen [19] design a multilevel model to predict the severity of an incident given the involved casualties (level 1), their respective vehicles (level 2) and the accident location (level 3). Ahmed *et al.* [20] conceive a multi-level model to predict crashes on a mountainous freeway by modeling both the dry or snow seasons and spatial correlation between adjacent sites. Deublein *et al.* [21] propose a multilevel model to manage simultaneously 4 response variables (resp. injury accidents, light injuries, severe injuries and fatalities) through gamma updating of the parameters. Finally, Fawcett *et al.* [22] predict future safety hotspots with a multilevel model where, first, the variance of the rate of a NB increases with the timestamp of an observation

[a]Gelman indicates in [18] that a half-normal with high standard deviation is a non-informative but proper prior for the variance parameter σ.

and, second, a global trend effect is altered by site-specific ones.

B. Nonparametric Machine Learning algorithms

Since the relationship between crash-frequencies and explanatory variables can be non-linear [3], ML algorithms often give more accurate predictions than do statistical models. In particular, this is the case for neural networks. Abdelwahab and Abdel-Aty [23] obtain better results with a back-propagation neural network (BPNN) than with an ordered logit statistical model to predict the severity of accidents at intersections. Chang [24] compares a one hidden layer BPNN with NB regression for crash frequencies prediction. The BPNN slightly outperforms the NB statistical model with a difference of 0.6% of accuracy on testing data. Huang *et al.* [25] compare a radial basis functions neural network (RBFNN) with BPNN and NB regression for crash frequencies prediction. The best results are obtained by RBFNN. In the experiments of Xie *et al.* [26] on crash-count prediction, a bayesian neural network outperforms both BPNN and NB regression. Similarly, Support Vector Machine (SVM) are used for crash-count prediction [27], crash-severity prediction [28], and the study of the effects on crash-count predictions of spatial correlations at different scales [29].

In all of these works, authors perform a sensitivity analysis of the black-box model: for each explanatory variable, while keeping all other variables unchanged, they record the effect on the output prediction of a perturbation of the current variable. However, as stated in [26], the relationship between the current explanatory variable and crash frequency may vary due to correlated variables, making it difficult to interpret the sensitivity plots. Moreover, by generating simulated data while assuming the explanatory variables to be independent, sensitivity analysis will indiscriminately generate potentially misleading hypothetical predictions for unlikely data points. Finally, this method only provides global interpretations at the scale of the whole network. It does not allow domain experts to identify roadway segments where the predictive model behaves singularly.

Furthermore, Karlaftis and Golias [30] as well as Chang and Chen [31] experiment with the CART algorithm for tree-based regression of crash-count data. The learned tree structure is effective for prediction (e.g. in [31], on a test dataset, tree-based regression obtains 52.6% accuracy against 52.3% for NB regression). However, it is designed to use few node-splitter variables (with the possible repetition of important ones) and key variables can be missing due to their correlations with sequences of node-splitters. Therefore, the decision tree itself cannot be used to explain the contributions of risk factors. Even if, by keeping track of candidate node splits during the tree-growing process, variables can be ordered by their predictive importance, this ordering is not visible in the final model and the measure of importance is not associated with confidence intervals. Moreover, according to [31], tree-based regression does not handle correctly interactions between risk factors.

To conclude, ML algorithms, being nonparametric, can benefit from large amount of data to cover big hypothesis spaces. In that way, they often outperform statistical models for crash-count prediction. However, they act as black-boxes and generate complex decision functions that lack interpretability. Although sensitivity analysis can partially remedy this problem, it has important drawbacks (e.g. no local analysis, assumption of independence of explanatory variables…) and is not as satisfactory as the thorough possibilities of analysis offered to road safety experts by Bayesian inferred GLM.

C. Local explanation models

To counteract the limitations of the global model-agnostic methods and mitigate the black-box effect, local explanation models have been proposed. Lundberg and Lee [4] introduce a class of explanation models under the name of additive feature attribution methods. They unify existing approaches such as LIME [32], DeepLIFT [33], Shapley regression values [34], etc. A model g in this class is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

where $z' \in \{0,1\}^M$, $\phi_i \in \mathbb{R}$ and M the number of simplified input features.

Let h_x be an input mapping from simplified inputs to original inputs (such that $x = h_x(x')$), and f a black-box model. An additive feature attribution model tends to reach:

$$g(z') \approx f\left(h_x(z')\right) \text{ whenever } z' \approx x'.$$

In [4], the authors prove that, under a choice of simple desirable constraints, and given a simplified input mapping h_x, there is a unique optimal additive feature attribution model whose coefficients ϕ_i correspond to the Shapley value of feature i — a game theoretic concept measuring how much feature i contributes to the prediction.

In [4], Lundberg and Lee present KernelSHAP, a novel approximation method for which the simplified input mapping is a binary vector indicating a subset S of features. Assuming feature independence, and given an instance x, the prediction of the model f on a simplified input vector is obtained by fixing the features present in S to their values in x, and by integrating over the marginal distribution of the features not present in S:

$$f\left(h_x(z')\right) = E_{X_{\bar{S}}}[f(x)]$$

To enhance the interpretability of tree-based ML algorithms, Lundberg *et al.* [35] propose the TreeSHAP variant that is able to account for feature dependence by integrating over the conditional distribution of the features not present in S.

Moreover, Lundberg and Lee [4] prove that, given an appropriate weight function $\pi_x(z')$, for an instance x, the

Shapley values of the original features are the coefficients of the linear model g minimizing the following loss function:

$$L(f, g, \pi_x) = \sum_{z' \in Z} \left[f\left(h_x(z')\right) - g(z') \right]^2 \pi_x(z')$$

$$\text{With } \pi_x(z') = \frac{M-1}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

Where M is the number of simplified input features and $|z'|$ is the number of present features in instance z'. Thus, for each instance, the Shapley values of the features are the solutions of this linear problem.

Thanks to its advantages, some safety analysis studies use SHAP to investigate the impact of different features on arterial incident duration [36] or in the context of real-time accident detection [37]. So far, to the best of our knowledge, SHAP has not been applied to long term crash prediction models. Moreover, prior studies rely on this method to enhance the interpretation of black-box models whereas in our approach, to be presented in the next section, it becomes a provider of objective priors obtained through the extraction of information from available data.

III. Methodology

A. Overview

In this section, we describe how to extract, from data and without prior expert knowledge, the information necessary to build a hierarchical Bayesian model with first-order interactions between explanatory variables to efficiently solve regression or classification problems while preserving interpretability. First, we elucidate how Shapley values of the variables give rise to a clustering of the original observations that is likely to make sense in terms of the problem to be solved. This clustering then informs a multilevel Bayesian model. Second, we explain how a self-organized neural network reveals the most important interactions between explanatory variables. These interactions are then integrated to the functional form of the multilevel model.

B. Supervised learning of a latent structure

We run the SHAP local explanation model (cf. related works) on a trained black-box ML algorithm used to solve the problem at hand. Given an observation, the individual contributions of the explanatory variables – in other words their Shapley values, see eq. (1) – appear on a *forceplot* (see Fig. 1 for an example of an observation coming from our crash prediction dataset). This visualization authored by Lundberg *et al.* [38] represents the forces by which the variables shift the output away from or towards the overall expected value of the model. Lundberg *et al.* also propose to group all the observations' forceplots by similarity so as to reveal a structure inherent to the unknown decision function of the ML algorithm (see Fig. 2).

Inspired by this visualization, we propose to discover a meaningful latent structure through a hierarchical agglomerative clustering (with the Ward linkage criteria) of the observations described by the Shapley values of the variables. We estimate the optimal number of clusters by detecting the greatest increase in the squared Euclidean distance between clusters when their number decreases (see Thorndike [39] for the original presentation of this widely used method). Experts can still fine-tune this number with the help of a dendrogram to obtain clusters that better match with their domain knowledge. For road safety studies, a cartographic projection of the clusters helps confirm their merits. This is illustrated in the next section.

Finally, the clusters are discovered on a train dataset on which we also grow a decision tree classifier to associate each observation to its cluster based on the values of the original explanatory variables. We rely on this classifier to link each observation of the test dataset to a cluster.

C. Supervised learning of interactions between explanatory variables

When first-order interactions between explanatory variables are integrated into a GLM, the relationship between a variable and the target may depend on the value of another variable. This can lessen the gap in predictive power between Bayesian inferred GLM and ML algorithms inherently able to capture complex nonlinear relationships. Moreover, these simple interactions are interpretable while the potentially highly entangled ones discovered by ML algorithms will often remain inaccessible to human understanding and increase the risk of overfitting.

In our approach, important first-order interactions are discovered with a variant of the Group Method of Data Handling (GMDH) family of supervised algorithms. This GMDH algorithm is a self-organized multi-layered structure of nodes (see Fig. 3). Each node generates its output z by applying a linear function with a covariation term to a pair of inputs (x_1, x_2) taken among either the nodes of the previous layer or the original explanatory variables:

$$z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2$$

Let n be the number of nodes of the previous layer and m be the number of explanatory variables. To build the next layer, for each of the $\binom{n+m}{2}$ polynomials, the $a_0 \dots a_3$ parameters are set by minimizing through Ridge regression the least square error made by the polynomial when it approximates the target on a train dataset. Then, the fitted polynomials are evaluated on a validation dataset to select the top m constituting the new layer. When the score obtained on the validation dataset by the best node of the last layer added stops improving, the process terminates and the best node of the penultimate layer is the output of the network.

Thus, this self-organized network discovers a polynomial that approximates the relationship observed on the training dataset between the explanatory variables and the target. In our approach, we use this polynomial to select the most important first-order interactions between explanatory variables. If the coefficient of a term involving the product of two variables exceeds a given percentage of the magnitude of the biggest

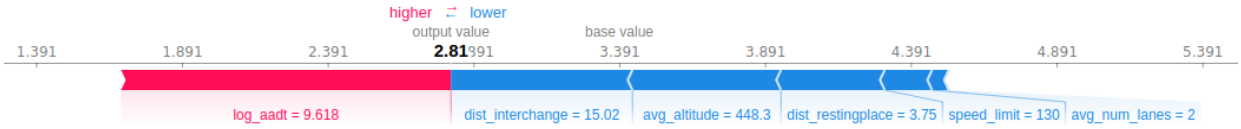


Fig. 1. Forceplot showing the variables' influence on the estimated crash count for a single observation.

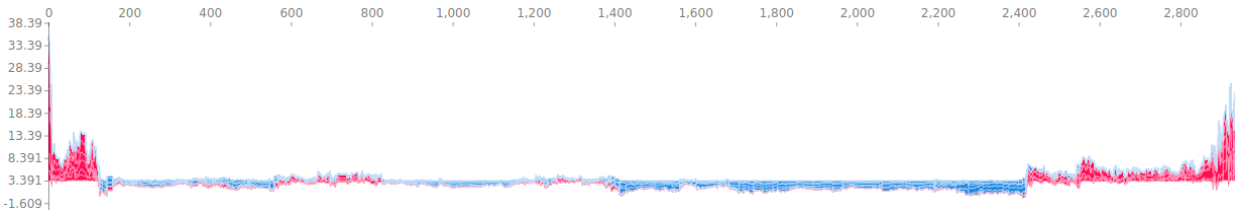


Fig. 2. Observations grouped by similarity of their forceplots.

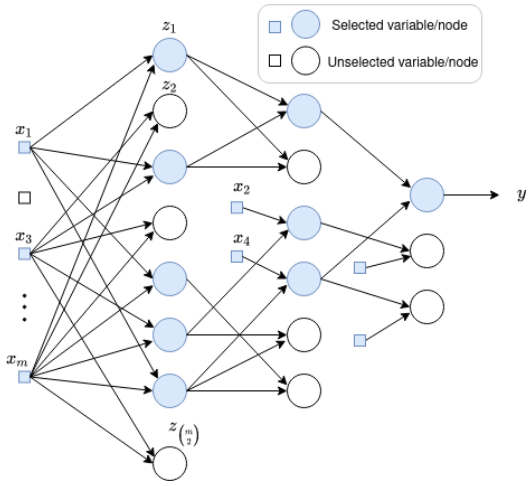


Fig. 3. Structure of the GMDH model.

coefficient, we add to our Bayesian hierarchical model an interaction between these two variables.

D. Bayesian Hierarchical model

a) Multilevel structure: To integrate the discovered latent structure, made of k clusters, into a GLM, we design a multilevel model:

$$
\begin{aligned}
Y_{ik} &\sim \mathcal{L}(Y_{ik}|\mu_{ik}, \Omega) \\
\mu_{ik} &= E[Y_{ik}|\mathbf{X}_{ik}] = g^{-1}(\eta_{ik}) \\
\eta_{ik} &= \beta_{0k} + \sum_{j=1}^{N} \beta_{jk} X_{ijk} \\
\beta_{jk} &\sim Normal(\mu, \sigma) \\
\mu &\sim Normal(0, 100) \\
\sigma &\sim HalfNormal(5)
\end{aligned}
$$

According to this model, output Y_{ik}, for observation i in cluster k, is generated from a likelihood distribution $\mathcal{L}$ parameterized with a set of specific parameters Ω and also μ_{ik}, the expected value of Y_{ik} conditioned on the observations. In the previous section, we saw that, for crash prediction, $\mathcal{L}$ can be a $NB(Y_{ik}|\lambda, \alpha)$ where λ is the expected number of crashes and α controls the amount of authorized over-dispersion. Next, η_{ik} is a linear transformation of the explanatory variables. The inverse link function g^{-1} is necessary to map the domain of η_{ik} (viz. the real line) to the one of μ_{ik}. For example, since the rate λ of a negative binomial must be positive, the exponential function is used for crash-count prediction. In case of binary classification, when the $Bernouilli(p)$ likelihood is used, p being a probability, g can be set to the logit function. Indeed, the logit maps a parameter constrained between 0 and 1 onto the real line (the inverse link function g^{-1} is, in that case, the logistic function). Finally, as explained in the previous section, the cluster-specific coefficients β_{jk}, for the N explanatory variables, depend on hyperpriors μ and σ, thus allowing an adaptive shrinkage to a mean common to all the observations.

b) Integration of interactions into the GLM: For a given observation i, let $\{int_{i1}, int_{i2}, \dots int_{iM}\}$ be the set of first-order interactions selected by our GMDH-based methodology. To integrate them into the Bayesian hierarchical model, just modify the linear form:

$$\eta_{ik} = \beta_{0k} + \sum_{j=1}^{N} \beta_{jk} X_{ijk} + \sum_{m=1}^{M} \beta_{mk} \text{int}_{im}$$

c) Model evaluation and validation: To use our model for point-estimate prediction, we must derive a Bayesian estimator from the posterior distributions. We use the mean of the posteriors which can be shown to minimize the mean squared error. In that way, we can compare our approach to blackbox ML algorithms with standard quality metrics on a test dataset. Moreover, to check if the estimation of the posterior distributions converged well, we use the Posterior Predictive Check (PPC) graphical analysis method. Indeed, according to Gelman and Hill in [40, p.158], PPC is a technique to "simulate replicated data under the fitted model and then compare these to the observed data". It allows one to

look for systematic discrepancies between real and simulated data [41].

IV. Experiments

In this section, we start by introducing the datasets used to compare our proposal against a state of the art black-box ML algorithm. Then, we describe our experimental setup, and we interpret the performances of the models. Finally, for the crash-count prediction dataset, we provide in-depth analysis of the key elements of our framework.

A. Datasets and preprocessing

The AREA dataset focuses on the forecasting of accidental highway segments located in the French Rhône-Alpes region. The network is divided into 430 2 km-long segments that cover both directions. After removing missing data (about 13% of the original data, mainly due to absent traffic), there remains $3,670$ observations for a 10 years period from January 2008 to December 2017. Table I describes the variables.

Furthermore, we conduct additional experiments on three public datasets:

- *AirBnB NYC dataset* [42] concerns the prediction of AirBnb housing prices in the New Yort City area, given parameters about the housing location, type and the number of reviews.
- *Breast cancer dataset* [43] focuses on a binary classification that determines whether a breast tumor is malignant or not, considering some parameters computed from a digitized image of a breast mass and characterizing the cell nuclei.
- *Insurance dataset* [44] concentrates on the prediction of the individual medical costs billed by health insurance, given parameters describing individuals (smoker, age, gender, etc.).

In preprocessing, categorical variables are one-hot-encoded and continuous variables are standardized.

B. Experimental setup

Our approach requires choosing an efficient ML algorithm whose predictions are analyzed in terms of Shapley values to extract a latent structure. We take the `XGBoost` [45] implementation of the gradient boosting tree algorithm which is often among the top performers of ML competitions such as Kaggle[b]. Its hyperparameters are optimized by grid-search. Shapley values are computed with an open source implementation[c] of the `TreeSHAP` algorithm [35]. The `scikit-learn`[d] implementation of the hierarchical agglomerative clustering algorithm is applied to the observations described by the Shapley values of the explanatory variables in order to materialize the latent structure. Then, to associate a new observation with its cluster, we train a decision tree classifier also from the `scikit-learn` toolkit. Moreover, to discover first-order interactions, we use `GmdhPy`[e], an open source Python implementation of the GMDH algorithm. Besides,Table II summarizes the values of the parameters and priors for the GLM on the four datasets. Bayesian inference is done with the *No U-Turn Sampler* (NUTS) [46] algorithm. Two Markov chains run for 3000 iterations with a burn-in period of 1000 iterations. The CPU-based `PyMC3` [47] library, with its efficient set of utilities, is employed for all datasets except Airbnb NYC for which the GPU-based `NumPyro` [48] (about 7 times faster in our experiments) is preferred due to the large number of observations.

[b]https://www.kaggle.com/

[c]https://github.com/slundberg/shap

[d]https://scikit-learn.org/stable/

C. Performance metrics

For regression tasks, we select the Root Mean Square Error (RMSE) and Mean Absolute Deviation (MAD) metrics:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

with n the number of observations, y_i the target and $\hat{y}_i$ the predicted value.

For the classification task, we use accuracy, recall and specificity:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{recall} = \frac{TP}{TP + FN}$$
$$\text{specificity} = \frac{TN}{TN + FP}$$

where TP, TN, FP and FN are respectively the numbers of true positives, true negatives, false positives and false negatives.

D. Results

Results summarized in Table III are averages of a 5-folds cross-validation.

a) AREA dataset: The so-called *local model* obtains the best results. Let us describe its strategy. With our 10-years-deep dataset, there are multiple observations for a same highway segment. Moreover, most explanatory variables associated with a segment, except the traffic, are constant over time. Therefore, for a segment in the test set, the local model predicts the average of the crash-counts observed for this same segment in the training set. Intrigued by the similar performances achieved by XGBoost and the local model, we discovered that, for nearby observations in the parameters' space, the paths taken through the XGBoost tree lead almost to the same leaves. Thus, on this dataset, the XGBoost algorithm tends to approximate the behavior of a naïve local model. So, these two best performing models tell us that a good strategy to predict a crash-count at a given location is to average the crash-counts already known to have occurred there. In this way, they are unable to provide information on risk factors correlated with observed accidents. In comparison,

[e]https://github.com/kvoyager/GmdhPy

TABLE I
DESCRIPTIVE STATISTICS OF THE DEPENDENT VARIABLE AND EXPLANATORY VARIABLES FOR THE AREA DATASET

Variables	Description	Mean	Std	Min	25%	50%	75%	Max
explanatory variables								
log_aadt	logarithmic value of the Average Annual Daily Traffic (AADT)	9.6	0.6	7.8	9.3	9.7	9.9	10.7
avg_num_lanes	Average number of lanes	2.2	0.5	1.0	2.0	2.0	2.0	4.0
avg_altitude	Average altitude (m)	332.3	152.8	92.4	226.6	268.8	400.7	890.6
avg_right_shoulder	Average rightshoulder width (m)	3.0	0.1	2.1	3.0	3.0	3.0	3.0
dist_interchange	distance to closest interchange (km)	3.3	3.8	0.0	1.0	2.3	4.3	25.0
dist_restingplace	distance to closest restingplace (km)	4.3	3.4	0.0	1.6	3.6	5.9	18.0
avg_speed_limit	average speed limit (km/h)	124.6	10.8	74.6	125.0	130.0	130.0	130.0
Dependent variable								
acc_tot	Traffic crash count (all severities considered)	3.42	4.0	0.0	1.0	2.0	4.0	60.0

TABLE II
PARAMETERS AND PRIORS FOR THE BAYESIAN HIERARCHICAL MODELS

Dataset	Link func.	N clusters	1st level	2nd level	3rd level
AREA	Log	4	$Y_{ik} \sim \text{NB}(\mu_{ik}, \alpha)^{\text{a}}, \quad \alpha \sim G(a = 0.5, b = 2.5)^{\text{b}}$	$\beta_j \sim \mathcal{N}(\mu, \sigma)$	$\mu \sim \mathcal{N}(0, 10), \quad \sigma \sim \mathcal{H}(5)$
AirBnb NYC	Identity	4	$Y_{ik} \sim \mathcal{N}(\mu_{ik}, 1000)$	$\beta_j \sim \mathcal{N}(\mu, \sigma)$	$\mu \sim \mathcal{N}(0, 100), \quad \sigma \sim \mathcal{H}(5)$
Insurance	Identity	2	$Y_{ik} \sim \mathcal{N}(\mu_{ik}, 1000)$	$\beta_j \sim \mathcal{N}(\mu, \sigma)$	$\mu \sim \mathcal{N}(0, 100), \quad \sigma \sim \mathcal{H}(5)$
Breast cancer	Logit	2	$Y_{ik} \sim \mathcal{B}\Big(p = \frac{\exp(\mu_{ik})}{1+\exp(\mu_{ik})}\Big)^{\text{c}}$	$\beta_j \sim \mathcal{N}(\mu, \sigma)$	$\mu \sim \mathcal{N}(0, 100), \quad \sigma \sim \mathcal{H}(5)$

[a] Negative Binomial distribution; [b] Gamma distribution; [c]Bernoulli distribution

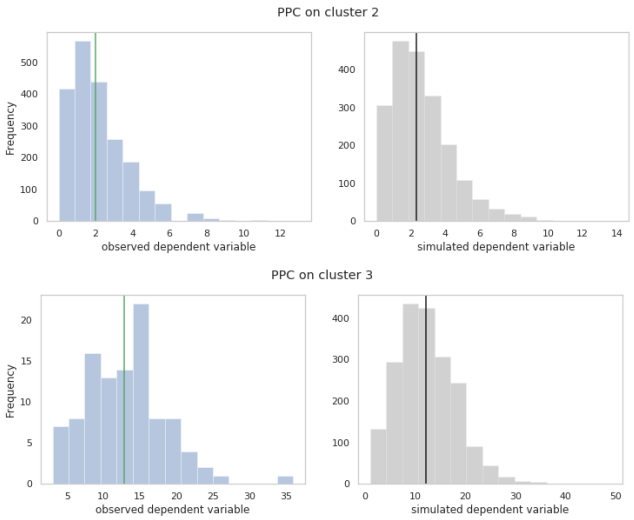


Fig. 4. Posterior Predictive Checks (PPC) on two clusters. Left column: observed data (cluster 2: 2064 samples; cluster 3: 104 samples). Right column: replicated data (2000 simulations).

our approach, although inferior in predictive performance to these two models, obtains a 17% increase in RMSE when compared to a standard Bayesian model. Moreover, we show in section IV-E that this improvement in predictive performance comes with enhanced interpretability.

In addition to out-of-sample evaluation, we validate our hierarchical Bayesian model with graphical PPC analysis. In Fig. 4, we compare, for two examples of clusters, the histogram of observed crashes with that of samples drawn from the posterior distribution of crash-counts. Green and black vertical lines indicate the means of, respectively, the observed and replicated data. The two distributions being similar, our model fits adequately the data. Thus, the integration of latent structure and interactions do not bring skewed prior knowledge that would disturb the inference.

Finally, to ease the validation of our results by AREA safety experts, we perform Hot Spot Identification (HSID), an essential tool for resources' allocation in safety management. We use the empirical Bayes (EB) method to identify 26 hotspots in 2017. Interested readers may refer to [49] for more details on HSID.

b) Other datasets: We observe the superior predictive performance of XGBoost, probably due to its ability to represent complex non-linear functions. This is also why this model is well-suited for the Shapley values' analysis that discovers a latent structure in connection with the predictive task. Indeed, our approach outperforms by a wide margin a standard Bayesian model. Finally, if the integration of first-order interactions improves only to a small extent the performances of the hierarchical model, it has nonetheless the potential to greatly enhance the quality of its interpretation, as demonstrated in the next section.

E. Analysis for the crash-prediction dataset

a) Latent structure: to analyze the clusters automatically discovered by our approach, we associate a map view from QGIS[f] (Fig. 5) with descriptive statistics (table IV). Clusters 0 and 3 are made up of highway segments located in the plains, close to major cities and with high traffic. Cluster 2, the largest, mainly comprised of 2-lanes segments in the

[f]https://qgis.org/

TABLE III
MODEL PERFORMANCE COMPARISON

Dataset	RMSE	MAD	
AREA			
Local[a]	2.73	1.77	
XGBoost	2.75	1.80	
Bayesian[b]	3.72	2.30	
BH[c]	3.11	2.02	
BH-int[d]	3.08	2.00	
AirBnB NYC			
XGBoost	0.434	0.306	
Bayesian	0.513	0.371	
BH	0.476	0.350	
BH-int	0.473	0.348	
Insurance			
XGBoost	4562	2600	
Bayesian	6119	4220	
BH	4920	2978	
BH-int	4912	2972	
	Accuracy	**Specificity**	**Recall**
Breast cancer			
XGBoost	0.96	0.94	0.98
Bayesian	0.95	0.92	0.97
BH	0.97	0.95	0.97
BH-int	0.97	0.94	0.97

[a] Local model averaging observed crash counts of closest neighbors
[b] Standard Bayesian inferred GLM
[c] Bayesian Hierarchical model without interaction effect
[d] Bayesian Hierarchical model with interaction effects

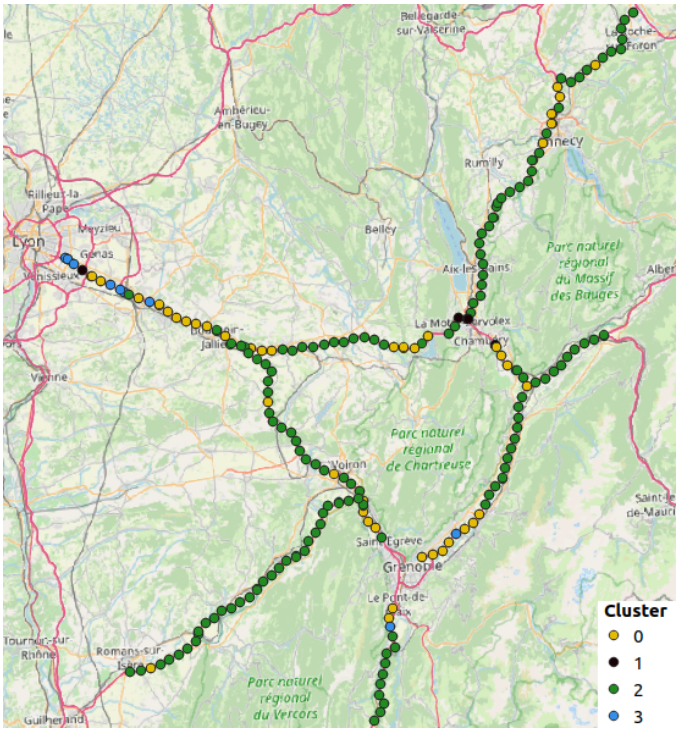
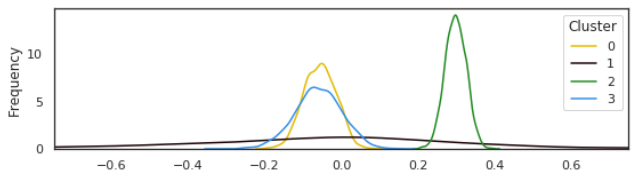
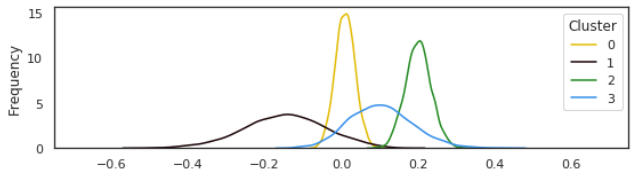


Fig. 5. Computed clusters of SHAP instances (2017) — each point represents the initial reference point of a 2 km-long section. For visual purposes, sections are only displayed for a single driving direction.



(a) log(AADT)



(b) speed limit

Fig. 6. Two examples of posterior distributions.

undulating countryside, counts the lowest number of accidents. Conversely, segments of cluster 1, the smallest, have the highest numbers of accidents. They are close to main interchanges near the major cities of Lyon and Chambéry. Moreover, those near Chambéry include a series of turns leading to a tunnel. This combination of potential risk factors could explain a large number of accidents. Thereupon, to validate this assumption, we are collecting data related to sinuosity and tunnels' positions to integrate them into our model. Note that pointing out such a small cluster made of singular accidental segments is of great value for the safety analysis process.

Furthermore, road safety experts from AREA confirm the relevance of these automatically discovered clusters. They point out that the time saved can advantageously be spent on, for example, planning remedial actions.

b) Bayesian hierarchical GLM: the analysis of the posterior distributions (Fig. 6) allows one to measure the influence of each variable on the crash count. Due to the inherent partial pooling nature of the model, posterior distributions are dissimilar among clusters thus revealing various impacts of the same explanatory variable on the crash count. For instance, the posterior's mean related to speed limit in Fig. 6b is positive for cluster 2 and 3, but negative for cluster 1.

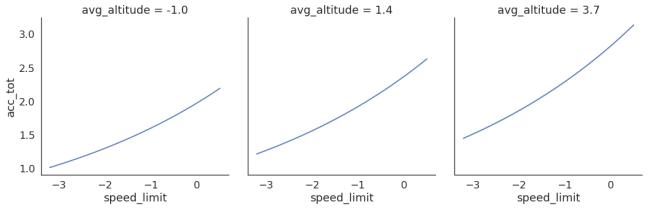
Moreover, we observe that the shapes of the posterior distributions are different from one cluster to another. This variability is linked to the size of the clusters: in general, the more samples, the more confidence in the estimates. For example, in Fig. 6a the posterior of $log(\text{AADT})$ has a sharpened probability distribution for cluster 2 but a flat one for cluster 1. The latter highlights a great uncertainty in estimating the coefficient associated with traffic and calls for extra vigilance when drawing conclusions for this risk factor.

c) Interaction effects: in our experiments, the GMDH polynomial highlights a major first-order interaction between speed limit and altitude. Triptych plots, introduced by Mc Elreath in [50, p.234], are made to visualize such interactions. Thus, Fig. 7 depicts the bivariate relationship between speed limit and predicted crash counts for cluster 2, depending on whether or not an interaction with the altitude is integrated into the Bayesian model. We observe that the slopes of the regression lines, constant and positive for both models, are

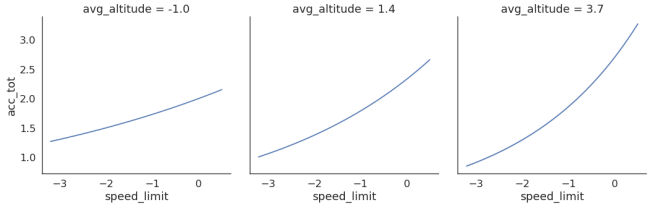
TABLE IV
DESCRIPTIVE STATISTICS OF THE DEPENDENT VARIABLE AND EXPLANATORY VARIABLES FOR EACH CLUSTER

Cluster	**0 (724[a])**				**1 (44)**				**2 (2064)**				**3 (104)**			
Variables	***mean***	***std.***	***min***	***max***	***mean***	***std.***	***min***	***max***	***mean***	***std.***	***min***	***max***	***mean***	***std.***	***min***	***max***
log_aadt	9.98	0.43	8.17	10.7	9.54	0.06	9.41	9.66	9.47	0.51	7.84	10.5	9.77	0.83	8.36	10.6
avg_num_lanes	2.49	0.59	2	4	2.09	0.20	2	2.5	2.04	0.24	1	3.8	3.17	0.64	2	4
avg_altitude	305	136	92.4	721	251	11.5	229	261	347	159	181	891	233	29.4	193	289
avg_right_shoulder	2.99	0.07	2.5	3	2.96	0.05	2.9	3	2.98	0.09	2.1	3	3	0	3	3
dist_interchange	2.37	3.18	0	23.0	1.5	1.14	0.59	3.75	3.62	3.90	0.07	25.0	2.08	1.87	0	5.34
dist_restingplace	4.43	3.81	0	18.0	8.12	3.82	0.15	10.3	4.09	3.15	0	15.0	5.01	3.53	0.15	14.0
speed_limit	122	10.2	90	130	85.8	9.68	74.6	102	127	8.15	90	130	113	16.6	90	130
acc_tot	5.39	2.99	0	18	19.0	11.9	2	60	2.01	1.80	0	13	12.8	5.44	3	36

[a]Number of samples for each cluster



(a) without interaction effect



(b) with interaction effects

Fig. 7. Triptych plots of predicted crash counts vs. speed limit on cluster 2.

steeper when considering an interaction. Thus, taking into account its interaction with altitude, the positive influence of speed limit on predicted crash counts gets more pronounced with high elevations.

V. CONCLUSION

Motivated by a crash-count prediction problem for a french highway network, we introduced a framework to build efficient and interpretable Bayesian hierarchical models for regression or classification tasks. Our main contribution is based on the data-driven discovery of objective priors in the form of a latent structure and strong first-order interactions between explanatory variables. We start with a trained and usually efficient, albeit opaque, ML algorithm in order to compute for each observation the Shapley values of the explanatory variables. Then, a latent structure, related to how the ML algorithm predicts the target, emerges from the hierarchical agglomerative clustering of the observations described by the Shapley values. Furthermore, we analyze the structure of a trained self-adaptive polynomial network to discover important first-order interactions.

Our experiments, conducted on four datasets, show that the integration of the latent structure and interactions into a Bayesian hierarchical model significantly improves the predictive performance compared to a traditional GLM. Moreover, while our model is less efficient than a state of the art black-box ML algorithm, it offers a high degree of interpretability. Indeed, the hierarchical structure allows a cluster-specific analysis of the posteriors with the possibility of quantifying the uncertainty associated with the coefficients of the explanatory variables. Interactions, on the other hand, tend to deservedly amplify the influence of key explanatory variables by bringing out configurations where their relationships with the target depend on secondary, context carrier, variables. Regarding highway safety, this enhanced interpretability helps experts in the field to accurately assess the risk factors and thus lead to appropriate policy decisions.

Finally, future works will focus on the adaptation of our framework to a temporal granularity gradually reduced until it approaches real time crash risk assessment. It will require the addition of explanatory variables (e.g. weather conditions, holidays, etc.) that could increase the advantage of ML algorithms able to produce complex non-linear decision functions. In this context, the challenge could even be greater to balance the quality of the predictions and their interpretability.

ACKNOWLEDGMENT

This work was supported by the French Association for Research and Technology (ANRT). We would like to thank APRR group, parent entity of AREA, for giving us access to their data and for sharing with us their expertise in road safety.

REFERENCES

[1] World Health Organization. (2020 (accessed July 16, 2020)) Road traffic injuries. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] French Road Safety Observatory. (2018 (accessed July 16, 2020)) Road safety annual report. [Online]. Available: https://www.onisr.securite-routiere.gouv.fr/sites/default/files/2019-09/Bilan%20de%20l%20accidentalit%C3%A9%20routi%C3%A8re%20de%20l%20ann%C3%A9e%202018%282%29.pdf

[3] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation research part A: policy and practice*, vol. 44, no. 5, pp. 291–305, 2010.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[5] B. Jones, L. Janssen, and F. Mannering, "Analysis of the frequency and duration of freeway accidents in seattle," *Accident Analysis & Prevention*, vol. 23, no. 4, pp. 239–255, 1991.

[6] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using linear and poisson regression models," *Transportation planning and Technology*, vol. 15, no. 1, pp. 41–58, 1990.

[7] S.-P. Miaou, "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions," *Accident Analysis & Prevention*, vol. 26, no. 4, pp. 471–482, 1994.

[8] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[9] W. Li, A. Carriquiry, M. Pawlovich, and T. Welch, "The choice of statistical models in road safety countermeasure effectiveness studies in iowa," *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1531–1542, 2008.

[10] X. Ma, S. Chen, and F. Chen, "Multivariate space-time modeling of crash frequencies by injury severity levels," *Analytic Methods in Accident Research*, vol. 15, pp. 29–40, 2017.

[11] P. Xu, H. Huang, N. Dong, and S. Wong, "Revisiting crash spatial heterogeneity: a bayesian spatially varying coefficients approach," *Accident Analysis & Prevention*, vol. 98, pp. 330–337, 2017.

[12] C. Walck *et al.*, "Hand-book on statistical distributions for experimentalists," *University of Stockholm*, vol. 10, 2007.

[13] American Association of State Highway Transportation Professionals (AASHTO). (2010) The highway safety manual. [Online]. Available: http://www.highwaysafetymanual.org

[14] S.-P. Miaou and D. Lord, "Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods," *Transportation Research Record*, vol. 1840, no. 1, pp. 31–40, 2003.

[15] D. Lord, A. Manar, and A. Vizioli, "Modeling crash-flow-density and crash-flow-v/c ratio relationships for rural and urban freeway segments," *Accident Analysis & Prevention*, vol. 37, no. 1, pp. 185–199, 2005.

[16] K. El-Basyouny and T. Sayed, "Comparison of two negative binomial regression techniques in developing accident prediction models," *Transportation Research Record*, vol. 1950, no. 1, pp. 9–16, 2006.

[17] D. Lord and P.-F. Kuo, "Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures," *Accident Analysis & Prevention*, vol. 47, pp. 52–63, 2012.

[18] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[19] A. P. Jones and S. H. Jørgensen, "The use of multilevel models for the prediction of road accident outcomes," *Accident Analysis & Prevention*, vol. 35, no. 1, pp. 59–69, 2003.

[20] M. Ahmed, H. Huang, M. Abdel-Aty, and B. Guevara, "Exploring a bayesian hierarchical approach for developing safety performance functions for a mountainous freeway," *Accident Analysis & Prevention*, vol. 43, no. 4, pp. 1581–1589, 2011.

[21] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber, "Prediction of road accidents: A bayesian hierarchical approach," *Accident Analysis & Prevention*, vol. 51, pp. 274–291, 2013.

[22] L. Fawcett, N. Thorpe, J. Matthews, and K. Kremer, "A novel bayesian hierarchical model for road safety hotspot prediction," *Accident Analysis & Prevention*, vol. 99, pp. 262–271, 2017.

[23] H. T. Abdelwahab and M. A. Abdel-Aty, "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections," *Transportation Research Record*, vol. 1746, no. 1, pp. 6–13, 2001.

[24] L.-Y. Chang, "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network," *Safety science*, vol. 43, no. 8, pp. 541–557, 2005.

[25] H. Huang, Q. Zeng, X. Pei, S. Wong, and P. Xu, "Predicting crash frequency using an optimised radial basis function neural network model," *Transportmetrica A: transport science*, vol. 12, no. 4, pp. 330–345, 2016.

[26] Y. Xie, D. Lord, and Y. Zhang, "Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 922–933, 2007.

[27] X. Li, D. Lord, Y. Zhang, and Y. Xie, "Predicting motor vehicle crashes using support vector machine models," *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1611–1618, 2008.

[28] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Analysis & Prevention*, vol. 45, pp. 478–486, 2012.

[29] N. Dong, H. Huang, and L. Zheng, "Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects," *Accident Analysis & Prevention*, vol. 82, pp. 192–198, 2015.

[30] M. G. Karlaftis and I. Golias, "Effects of road geometry and traffic volumes on rural roadway accident rates," *Accident Analysis & Prevention*, vol. 34, no. 3, pp. 357–365, 2002.

[31] L.-Y. Chang and W.-C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *Journal of safety research*, vol. 36, no. 4, pp. 365–375, 2005.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[33] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[34] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[35] S. Lundberg, G. Erion, H. Chen, A. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[36] A.-S. Mihaita, Z. Liu, C. Cai, and M.-A. Rizoiu, "Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting," *arXiv preprint arXiv:1905.12254*, 2019.

[37] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis," *Accident Analysis & Prevention*, vol. 136, p. 105405, 2020.

[38] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[39] R. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[40] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[41] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[42] AirBnb. NYC housing prices. Accessed August 18, 2020. [Online]. Available: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

[43] Breast cancer Wisconsin (diagnostic). Accessed August 18, 2020. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[44] Medical cost personal. Accessed August 18, 2020. [Online]. Available: https://www.kaggle.com/mirichoi0218/insurance

[45] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[46] M. D. Hoffman and A. Gelman, "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, 2014.

[47] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using pymc3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[48] D. Phan, N. Pradhan, and M. Jankowiak, "Composable effects for flexible and accelerated probabilistic programming in numpyro," *arXiv preprint arXiv:1912.11554*, 2019.

[49] A. Montella, "A comparative analysis of hotspot identification methods," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 571–581, 2010.

[50] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.