TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Freight Focus

Transportation Research Record
2023, Vol. 2677(2) 299–309
© National Academy of Sciences: Transportation Research Board 2022
Article reuse guidelines: sagepub.com/journals-permissions
DOI: 10.1177/03611981221133085
journals.sagepub.com/home/trr
SAGE

Ranking Risk Factors in Financial Losses From Railroad Incidents: A Machine Learning Approach

Neeraj Dhingra[1], Raj Bridgelall[2], Pan Lu[1], Joseph Szmerekovsky[1], and Bhavana Bhardwaj[3]

Abstract

The reported financial losses from railroad accidents since 2009 have been more than US$4.11 billion dollars. This considerable loss is a major concern for the industry, society, and the government. Therefore, identifying and ranking the factors that contribute to financial losses from railroad accidents would inform strategies to minimize them. To achieve that goal, this paper evaluates and compares the results of applying different non-parametric statistical and regression methods to 15 years of railroad Class I freight train accident data. The models compared are random forest, *k*-nearest neighbors, support vector machines, stochastic gradient boosting, extreme gradient boosting, and stepwise linear regression. The results indicate that these methods are all suitable for analyzing non-linear and heterogeneous railroad incident data. However, the extreme gradient boosting method provided the best performance. Therefore, the analysis used that model to identify and rank factors that contribute to financial losses, based on the gain percentage of the prediction accuracy. The number of derailed freight cars and the absence of territory signalization dominated as contributing factors in more than 57% and 20% of the accidents, respectively. Partial-dependence plots further explore the complex non-linear dependencies of each factor to better visualize and interpret the results.

Keywords

rail, rail safety, machine learning, human factors in crashes, system safety, data science, class I rail, freight train

Every year, railroads invest an average of 40% of their revenue on capital expenditures, maintenance, and condition monitoring (*1*). Despite those investments, the high number of accidents falls far short of the goal of the Federal Railroad Administration (FRA) (*2*) to reduce rail-related accidents, injuries, and fatalities to zero. For a decade before 2019, nearly 25,000 accidents caused 446 deaths, 5137 injuries, and more than US$4.11 billion in financial loss seasonally adjusted to 2018 dollars (*3*). Class I railroads accounted for 78% of those accidents, more than 72% of the resulting injuries and fatalities, and 81% of the total financial loss. Figure 1 summarizes the annual Class I railroad accidents and the financial losses for the decade before 2019.

The consistently large number of accidents and the injuries and fatalities they cause place a significant social and economic burden on the industry, the environment, and society. Therefore, it is vital to understand the dominant accident causes to guide strategies and policies that could minimize financial losses from accidents. Subsequently, the goal of this paper is to apply data mining (DM) and machine learning (ML) techniques to 15 years of Class I freight railroad accident data from 2004 to 2018 to reveal insights about the major factors contributing to financial losses from Class I freight train accidents.

FRA maintains historical data of railroad accidents in three primary databases. These datasets contain greater variety and have grown far beyond the ability of humans

[1] Department of Transportation, Logistics, and Finance, North Dakota State University, Fargo, ND
[2] Department of Transportation, Logistics and Finance, North Dakota State University, Plano, TX
[3] Department of Computer Systems and Software Engineering, Valley City State University, Valley City, ND

Corresponding Author:
Neeraj Dhingra, neeraj.dhingra@ndsu.edu

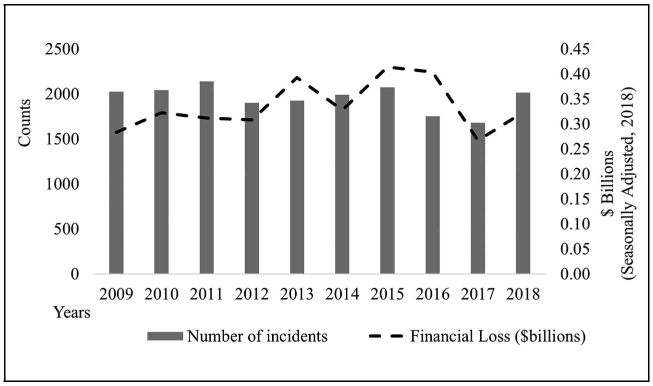

Figure 1. Class I railroad incidents from 2009 to 2018 and the reported financial loss.

and commonly used software tools to capture, manage, and process data within a "tolerable elapsed time" (*4*). The available accident data are in non-uniform formats. The data includes heterogeneity, variety, unstructured features, missing values, incorrectly formatted values, and redundancy (*5*, *6*). Therefore, it is not possible to apply standard statistical methods directly to the raw data. Therefore, advanced techniques such as DM and ML are necessary to prepare the data for processing.

DM is helpful in analyzing vast amounts of data by using many different techniques to discover useful patterns and relationships among features (*7*, *8*). Kohavi (*9*) specified that insight and prediction are the two primary goals of DM. Insights identify patterns and trends that are useful, whereas prediction leads to the identification of a model that provides reliable forecasts based on new input data. Many researchers have applied different DM/ML methodologies to analyze factors that cause accidents on roadways (*10–12*), at highway rail-grade crossings (HRGCs) (*13–18*), and on railways (*19*, *20*). For instance, Sohn and Lee (*12*) compared the results of neural network, Bayesian fusion, decision tree (DT), bagging, and clustering models on Korean road accident data. Their results indicate that clustering-based classification works better than the other methods. Depair et al. (*11*) also examined clustering techniques to identify homogenous accident types. They used vehicle types as the basis for segmentation and evaluated the relationship to injuries caused by different segments.

Some researchers used DM techniques to analyze road-related factors and linked them to accident severity. Beshah and Hill (*21*) compared different DM models to investigate the role of road-related factors in accident severity in Ethiopia and concluded that *k*-nearest neighbors (KNN) performed best. Mousa et al. (*22*) compared the ability of tree-based ensemble methods to predict the onset of lane changing maneuvers by using connected vehicle data and found that the extreme gradient boosting method (XGBM) performed best. The highest accuracy was 99.7%, and that was better than methods using DTs, gradient boosting (GB), and random forest (RF) ensemble methods.

Other related areas of research focused on HRGC accidents. Hu et al. (*16*) evaluated the relationship between crash frequency and the relevant attributes of highway and railroad systems. Ghomi et al. (*13*) used DM techniques to identify some of the main factors associated with the injury severity of road users involved in HRGC accidents. Kang and Khattak (*17*) investigated the severity of HRGC accidents by clustering the data using a combination of DM and statistical methods. Brown (*19*) applied text mining to identify factors contributing to railroad accidents. Mirabadi and Sharifian (*20*) used association rule mining to reveal the relationships and patterns in Iranian railway accident data. Many other researchers have conducted studies that use other analytical criteria to discover relationships between accident risk and contributing factors (*8*, *23–25*).

All research that analyzed rail or road accidents using DM techniques focused on identifying contributing factors that relate to attributes of the respective infrastructure. There is a gap in the research to identify and rank risk factors in financial loss from railroad accidents. Subsequently, the main contribution of this research is a comparison of the ability of different non-parametric, tree-based DM methods, and a regression model to identify the risk factors in financial loss by analyzing 15 years of railroad Class I freight train accident data. The authors then use the best predictive model to rank the major factors based on their influence on financial loss. This research extends previous work on railroad safety in the following two ways:

1. it isolates factors that lead to financial losses;
2. it ranks the importance of the major contributors.

The remainder of the paper is structured as follows: the next section introduces the models used to identify the factors that influence financial loss. The section that follows describes the data structure, variables, data cleaning, and data handling. After that, a section compares the model outputs for selection, variable ranking, and the marginal effect of the variables. The final section presents concluding remarks and describes future work.

Model Development

This study used tree-based models (RF, stochastic GB, and extreme GB), the *k*-nearest neighbor method, and the support vector machine (SVM) to classify the data according to the selected features or factors. In addition, stepwise linear regression (SLR) provided a baseline for comparison because of its proven effectiveness in previous research (*26*, *27*). The next sections provide basic descriptions of the six models, all available from the caret package of the R Project for Statistical Computing.

Model Regularization

Model regularization involves trading off training data bias for a reduced variance on new data. This is achievable by partitioning the data appropriately into development and test sets. The former is used for cross-validation while tuning the model, and the latter is used to test the final regularized and tuned model (*28*). Running the models with many different variations in partitioning revealed that a 70/30 split between development (training/validation) and testing datasets yielded the lowest variance.

k-Nearest Neighbors Method

KNN is a supervised learning algorithm that uses a non-parametric technique that does not require any assumptions on the underlying data distribution. This algorithm predicts the class of an observation by searching through the entire dataset to identify *k* other observations that are most like it, and then takes the class associated with the majority. The measure of similarity is based on one of several available distance measures (*29*, *30*). This analysis selects the Euclidean distance measure because it is the most common.

Random Forest

Standard DTs split the dataset by selecting an attribute and a threshold that maximizes the purity of the subtrees. The purity of a node increases as the class imbalance of the dataset within that node increases. However, this tree-splitting strategy results in trees that tend to over-fit the data and subsequently fail to regularize by exhibiting a high variance on new data. The RF addresses the regularization issue by introducing two levels of randomness—namely the random selection of learning data and the random selection of decision attributes for tree splitting. Such an adjustment results in better performance than many other classifiers models, and improves robustness against over-fitting (*31*, *32*).

The RF learns an ensemble of trees by bootstrapping the same dataset through random sampling with a replacement, and then randomly selecting a predetermined number of attributes for subsequent tree splitting (*32*). The selected class of observation is the majority vote from all trees created—also referred to as aggregation. Subsequently, the literature often refers to the combined methods of bootstrapping and aggregation as the bagging method. Bagging does not require tree pruning for regularization, because averaging the results of all bootstrapped samples reduces the variance (*33*).

Stochastic Gradient Boosting

The stochastic gradient boosting model (SGBM) is an extension of the GB technique. Gradient refers to model building optimization during the learning process. Boosting refer to finding a more accurate hypothesis by combining the predictions of many weak hypotheses (learners), each of which is moderately accurate (*34*). Most of the time, learners are non-linear models (decision or regression trees), and for such cases, the literature refers to GB as "gradient tree boosting" (GTB). The GTB algorithm builds an ensemble of weak prediction models by adding a sequence of trees, with successive trees grown on reweighted versions of the data. At each stage, GTB generates a new tree from the residuals and adds to the existing group of trees. The algorithm builds the final ensemble with a weighted summation of the individual learners.

Motivated by Breiman's bagging phenomenon, Friedman (*35*, *36*) augmented the GB procedure and incorporated randomness as part of the GB algorithm, calling the resulting technique the SGBM. Friedman recommended that instead of using the entire dataset to perform the boosting, it is more appropriate to select a random subsample from the training dataset at each step of the boosting process. The base learner then uses this randomly selected subsample.

Extreme Gradient Boosting

The XGBM extends the GB method for greater efficiency and accuracy. Unlike the GB technique, the XGBM implements an additional regularization to avoid over-fitting by imposing additional control over model complexity (*22*). The additional regularization term does not depend on the randomness. Instead, the focus of this additional term always remains on minimizing the model complexities based on some leaves and the sum-of-square scores of those leaves. For further reference, Bridgelall (*37*) presents a detailed study on the XGBM.

Support Vector Machine

A SVM is a non-parametric statistical learning technique that requires no assumption on the underlying data distribution. The concept is to separate data across a decision boundary (hyperplanes) determined by a small subset of the data (feature vectors). The data subset that supports the decision boundary is called the support vector (*38*). The SVM assumes that the multi-feature data are linearly separable in the input space. However, in practice, data points of different hyperplanes overlap, which makes linear separability challenging (*39*). A "kernel trick" overcomes the problem of the linearity restriction on the decision boundary. The kernel trick uses a transformation function to map the input vector into a higher dimension space by introducing new parameters (*38*). The "trick" part is that the SVM operates only on the vectors in their ambient space, without actually transforming the vector into a higher dimension. This analysis uses the radial kernel. Bridgelall (*37*), Mountrakis et al. (*38*), and Yoonsuh and Hu (*40*) explain the use of the kernel trick in more detail.

Stepwise Linear Regression

SLR is the process of building a model by successively removing or adding feature variables based on their relationship with the response variable. In other words, SLR is a method of regressing multiple variables in multiple stages. In each stage, the method removes or adds variables based on their correlation with the response variable.

Model Comparison

To minimize the potential for over-fitting or under-fitting, the ML procedure incorporates a *K*-fold cross-validation process with *N* repeats to identify the best model parameters. As explained by Jhangiri and Rakha (*40*), the *K*-fold algorithm segments the training data randomly into *K* parts or folds of approximately equal size. Subsequently, the algorithm builds a model from the union of the remaining *K*–1 folds and evaluates the model performance on the validation fold. The algorithm repeats the cross-validation *K* times so that each fold serves as the validation data exactly once. The algorithm repeats the *K*-fold process *N* times to introduce further randomization. The algorithm builds the final model by using those parameters that produce the best average performance across the *K* validations.

The *K*-fold cross-validation algorithm sets a uniform random seed before training each model to ensure consistency in the data partitions and repeats. Once trained, the process adds all the models to a list for re-sampling. This function verifies that the models are comparable and have used the same training scheme (*41*, *42*). Finally, the algorithm evaluates the performance of the models by comparing the mean absolute error (MAE), the root mean squared error (RMSE), the mean absolute percentage error (MAPE), and the *R*-squared metrics. The MAE is the unweighted average of the absolute differences between the predicted and actual observations. The RMSE is the square root of the average of the squared differences between the predicted and actual observations. The MAPE is the average of the absolute percentage of prediction errors. Therefore, the RMSE represents the average magnitude of the error and the MAPE represents the magnitude of percentage of the error relative to financial loss. *R*-squared is a measure of the percentage of the variation in the response variable that the model explains.

Data

FRA requires that railroads maintain and submit a detailed report of all significant accidents or incidents associated with railroad operations. FRA compiles these reports in the railway equipment accident (REA) database (*25*). This study used 15 years of REA accident data from all railroads reporting all types of accidents between 2004 and 2018 (*3*). This database records all accidents that exceed a specified financial cost (the inflation-adjusted 2019 threshold was US$10,700) from damages to on-track equipment, signals, tracks, track structures, and roadbeds (*43*). However, there are some other significant financial factors that are not considered while estimating the actual financial damage from a rail accident. Such expenses include delays, re-routing, emissions, cargo losses, first and emergency responders, and other operating costs. Those indirect expenses could add up to a significant amount and could be included in the actual financial damage. However, those indirect factors are often not reported or available. Subsequently, this study uses Class I freight train accident data for greater consistency in the analysis. The data consists of more than 145 variables, such as the railroad identifier, accident location, speed, and other attributes that attempt to describe the nature of the event. A limitation of this database is that it may not capture all the underlying factors that contributed to the level of financial loss. However, the models are based only on the available factors and are likely to expose dominant factors in causing financial loss.

Cleaning and Structuring

The data cleaning followed a three-step process. The first step deleted variables that were not appropriate, such as text narratives, dummy variables, and duplicate variables. The second stage removed variables such as "number of engineers" and "location" that did not support the analysis objectives. The third stage modified some of the FRA-structured default variables. Figure 2 presents a flow chart explaining the variable selection process. Restructuring of the default variables, also called feature engineering, was performed as follows.

(1) TIMEHR—changed the specific hours and minutes of the incident from the standard 12-h,

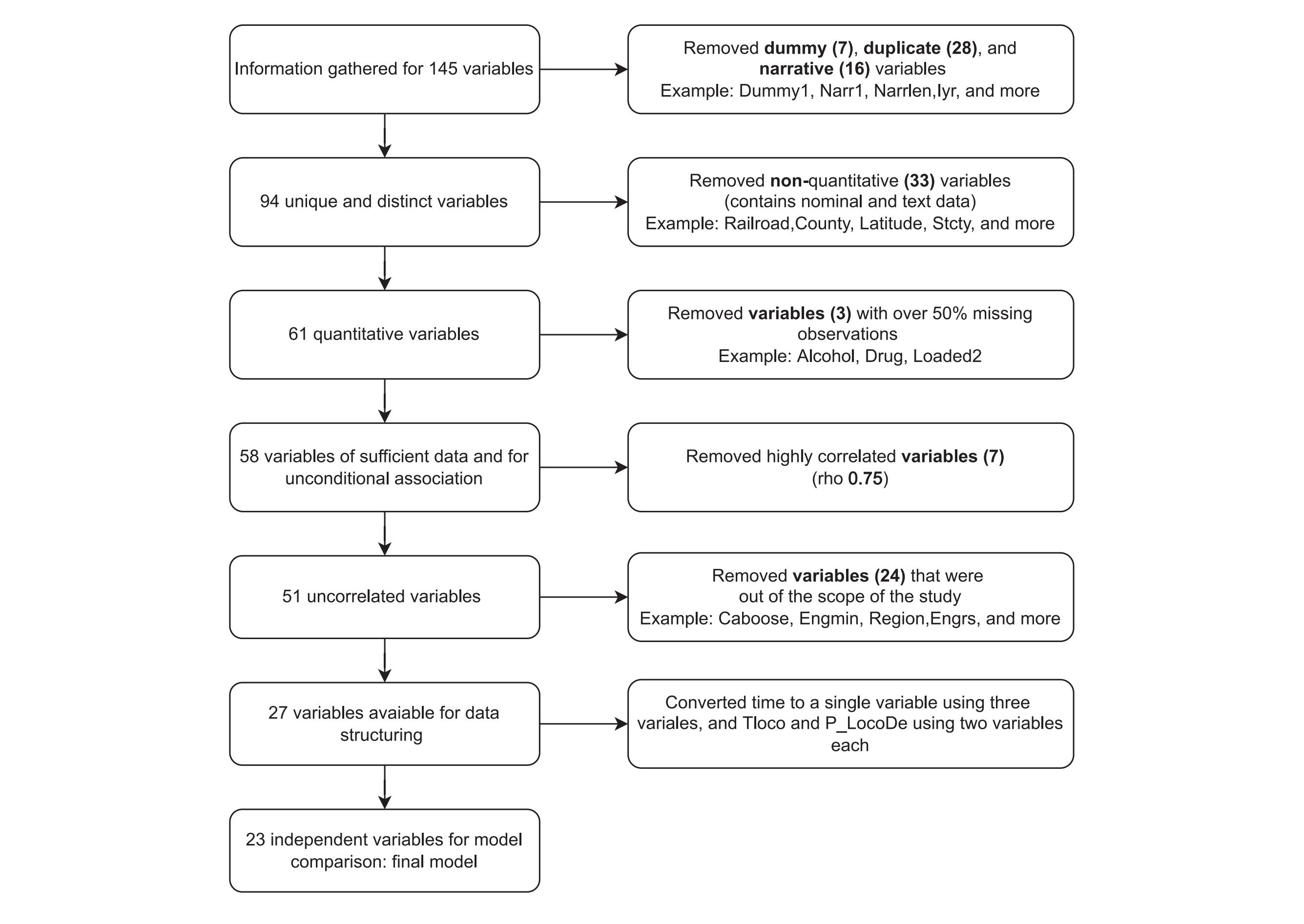
Information gathered for 145 variables
Removed **dummy (7)**, **duplicate (28)**, and **narrative (16)** variables
Example: Dummy1, Narr1, Narrlen,Iyr, and more
94 unique and distinct variables
Removed **non-**quantitative **(33)** variables
(contains nominal and text data)
Example: Railroad,County, Latitude, Stcty, and more
61 quantitative variables
Removed **variables (3)** with over 50% missing observations
Example: Alcohol, Drug, Loaded2
58 variables of sufficient data and for unconditional association
Removed highly correlated **variables (7)**
(rho **0.75**)
51 uncorrelated variables
Removed **variables (24)** that were out of the scope of the study
Example: Caboose, Engmin, Region,Engrs, and more
27 variables avaiable for data structuring
Converted time to a single variable using three variales, and Tloco and P_LocoDe using two variables each
23 independent variables for model comparison: final model

Figure 2. Variable selection flow chart.

a.m.–p.m. format to a single variable in 24-h military time format.

(2) P_CARSDMG—a new variable that is equal to the percentage of cars carrying hazmat that were damaged or derailed.

(3) TRKCLAS—changed the FRA track classes of A–E to a numeric categorical variable for compatibility across the DM techniques used.

(4) TRKDNSTY—imputed missing values and replaced zero values based on the maximum reported for that county.

(5) Ospeed—restructured “train speed” as a categorical variable over speed where the value is “1” if the train was traveling faster than the track class limit, and “0” otherwise.

(6) P_LocoDe—a new variable that contains the percentage of locomotives derailed is estimated using the same dataset.

(7) Tloco—a new variable that contains the total number of locomotives is obtained using the same dataset.

(8) Cause—changed the primary cause of an accident to a categorical variable with five classes based on their alphabetic order. “1” = mechanical and electrical failure”; “2” = miscellaneous causes not otherwise listed; “3” = rack, roadbed, and structures; “4” = signal and communication; and “5” = train operation—human factors.

(9) EQATT—“1” if someone was attending the equipment and “0” otherwise.

(10) R_ Amount—a modified dependent variable containing the total reported financial damage. The modifications are as follows.

   a. Time value normalization: adjusted the total reportable damage from the variable ACCDMG to the average consumer price index seasonally adjusted amount of 2018.
   b. The REA databases should include only those accidents that exceed financial losses of US$10,700. Therefore, this adjustment deleted records with lower amounts because such entries may be included in

Table 1. List of Variables and their Description

Variable	Description	Variable type
R_Amount	Seasonally adjusted financial loss based on 2018 prices (dependent variable)	Continuous
MONTH	Month of the incident	Categorical
DAY	Day of the incident	Categorical
TIME	Time of the accident (military standard time)	Continuous
TYPE	Type of accident (1–13)	Categorical
P_CARSDMG	% of hazmat cars damaged or derailed	Continuous
TEMP	Temperature in degrees Fahrenheit	Continuous
VISIBLTY	Daylight period (1–4)	Categorical
WEATHER	Weather conditions (1–6)	Categorical
Ospeed	Boolean of train traveling over the speed limit	Categorical
TONS	Gross tonnage, excluding power units	Continuous
EQATT	Boolean for equipment attended by a human	Categorical
TRKCLAS	FRA track class (0–9)	Categorical
TRKDNSTY	Annual track density—gross tonnage in millions	Continuous
POSITON1	Car position in train (first involved)	Categorical
POSITON2	Car position in train (causing)	Categorical
Tloco	Total number of locomotives	Categorical
P_LocoDe	Percent of locomotives derailed	Continuous
LOADF2	Number of derailed loaded freight cars	Categorical
EMPTYF2	Number of derailed empty freight cars	Categorical
CAUSE	Primary cause of incident	Categorical
TOTKLD	Total killed for the railroad as reported	Categorical
SIGNAL	Type of territory—signalization	Categorical

Note: FRA = Federal Railroad Administration.

error in the dataset and not represent most accidents.

c. Further analysis was conducted using the interquartile range (IQR) method to identify any outliers in the reported financial loss variables. The distribution revealed some variables within the 5-percentile and beyond the 95-percentile that were eliminated.

Handling Correlation and Missing Values

Missing values do not cause a problem for DT models because the method imputes those values based on the values of other observations that are in similar classes.

However, models such as linear regression (LR) cannot use data that contain missing values, thereby making the size of the dataset inconsistent for uniform comparison of models (*28*). Model comparison is most appropriate between models that are fitted using the same set of observations (*28*). Therefore, it is necessary to impute missing values before fitting models for comparison of performance. This analysis replaced missing values using an approach based on KNN, referred to as KnnImputation. The model identified "k" closest observations for each missing value based on the Euclidean distance and computed the weighted average as the missing value. Researchers observed that using $k = 10$ provided a good trade-off of low computational cost and low biases in the model estimates (*44*). Therefore, this study also uses $k = 10$ for imputing missing values

Highly correlated variables with the dependent variable are redundant and do not contribute additional information in the model (*45*). Therefore, the procedure removed those variables that had a correlation coefficient above a commonly selected threshold of 0.75 (*46*).

Dataset for Model Comparison

The final dataset contained 23 variables (Table 1) and approximately 12,500 observations of freight train accidents of Class I railroads.

Results and Discussion

Model Selection

Table 2 summarizes the evaluation metrics for the six ML models and their respective training times, using 10-fold cross-validations with three repeats. In general, the ensemble tree-based models outperformed the other models. Among tree-based ensemble methods, the XGBM provided the best predictive capability based on the lowest RMSE, MAE, and MAPE metrics, and the

Table 2. Model Comparison Evaluation

Models	Label	MAE	RMSE	R^2	MAPE	Model running time (h)
GBM	Gradient boosting model	87,131	139,295	0.46	32.74	7.6
KNN	*k*-nearest neighbors	122,391	189,069	0.03	45.99	2.4
SVM	Support vector machine	102,771	204,053	0.05	38.62	2.3
RF	Random forest	88,939	143,402	0.45	33.42	10.3
STEPWISE	Stepwise regression	95,392	149,052	0.40	35.84	1.5
XGBM	Extreme gradient boosting method	85,989	137,646	0.46	30.97	6.2

Note: MAE = mean absolute error; RMSE = root mean squared error; MAPE = mean absolute percentage error.

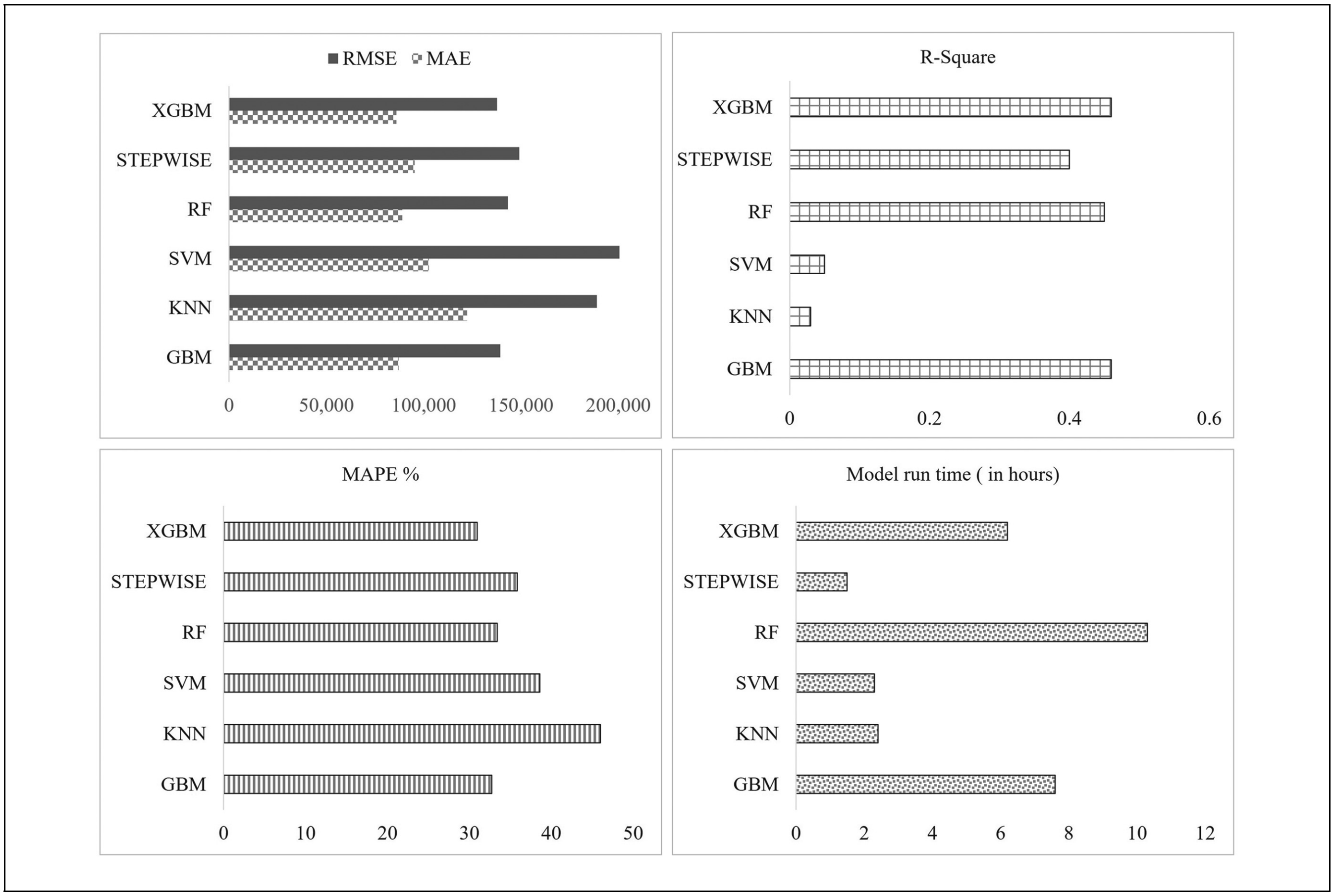

Figure 3. Mean absolute error (MAE), root mean squared error (RMSE), *R*-squared, and model training time (in hours).
Note: XGBM = extreme gradient boosting method; RF = random forest; SVM = support vector machine; KNN = *k*-nearest neighbors; GBM = gradient boosting model.

highest *R*-squared metric. On the other hand, tree-based models required the maximum time for training. Moreover, the RF required the longest time amongst all of the six models. Therefore, the final model was selected based on the model performance parameters and time required to train the models, which is the XGBM. Figure 3 provides a visualization of the MAE, RMSE, and *R*-squared for all six models.

Variable Importance using the XGBM

After identifying the XGBM as the best model for the data, the analysis focused on identifying the significant contributors to the prediction accuracy. Table 3 summarizes the results. The model ranked importance factors with regard to gain, which is a measure of the average gain in purity when splitting the training data for each tree of the model (*47*). Therefore, the gain is proportional

Table 3. Results of Variable Importance

Feature	Description	Gain	Frequency	Cover
LOADF2	# of derailed loaded freight cars	0.57459	0.2900	0.51493
SIGNAL1	Type of territory—signalization (mandatory)	0.20220	0.1700	0.05337
EMPTYF2	# of derailed empty freight cars	0.10124	0.1366	0.26967
TRKCLAS4	FRA track class 1–9	0.06536	0.1726	0.02682
TONS	Gross tonnage, excluding power units	0.02092	0.0757	0.06610
TRKCLAS3	FRA track class 1–9	0.01018	0.0460	0.00454
TRKCLAS2	FRA track class 1–9	0.01008	0.0320	0.00698
TYPE3	Type of accident: 03 = rear-end collision	0.00599	0.0197	0.02691
P_LocoDe	% of locomotive derailed	0.00370	0.0263	0.01545
CAUSE	Contributing cause of incident	0.00233	0.0091	0.00182
POSITON1	Car position in train (first involved)	0.00194	0.0089	0.01257
POSITON2	Car position in train (causing)	0.00069	0.0043	0.00037
TRKDNSTY	Annual track density—gross tonnage in millions	0.00062	0.0063	0.00033
MONTH12	Month of the incident	0.00014	0.0023	0.00012
Tloco	Total number of locomotives	0.00001	0.0003	0.00001

Note: FRA = Federal Railroad Administration.

to its importance in generating predictions. The cover and frequency provide additional indicators about the importance of those variables in building the model during training. Frequency is the percentage of time that the model used the corresponding feature to split the training data across all trees. The cover is the frequency weighted by the number of training data observations involved with those splits.

The results indicate that the number of loaded freight cars derailed is most strongly associated with financial losses from accidents by a proportional contribution of 57%. Territory signalization (SIGNAL) is the second most strongly associated factor by a proportional contribution of more than 20%. The number of empty freight cars derailed is next, which improves the predictability by more than 10%. Accidents on track Class 4 are the next factor most associated with financial losses by a proportional contribution of more than 6%. Table 3 summarizes the rank of the other variables.

Marginal Effect of Predictor Variables

Advanced ML models can significantly improve predictions and classifications, but understanding the influence of one or more predictor variables on the response variable is not feasible even with these advanced models. Partial-dependence plots (PDPs) can show the marginal effect of a single attribute on the predicted outcome of a ML model (*4*). The PDPs show the distinct impact of the most influential variables by marginalizing over the effects of all other variables in the model (*48*). The process starts with fitting the best performing ML model (the XGBM), followed by using the partial-dependence functions in the PDP package of R-studio with default parameter settings to visualize the complex non-linear global relationship between each factor and the predicted outcome.

Figure 4 shows that, except for the effects of the binary signal variable, the PDPs from the XGBM exhibit non-linear patterns. The yhat ($\hat{y}$) variable actually does not represent the predicted financial loss; instead, it represents the change in financial loss with the change in value of each predictor variable.

Per the results, financial damage generally increased with the number of derailed cars (LOADF2) and peaked at 40. Non-signaled territories (SIGNAL = 2) are associated with higher financial losses than with territories that are signaled. The partial dependency on EMPTYF2 suggests that financial losses tend to be most severe when 30–40 empty cars derail. Financial loss generally increases with track classification, and peaks for Class 7 tracks. Trains that carry approximately 20,000 tons tend to more significantly influence financial losses. Head-on collisions (TYPE2) and rear-end collisions (TYPE3) are associated with higher financial losses than other accident types. Accident cause (CAUSE) category 5 (human factor related) is associated with the highest financial losses. P_LocoDe (percentage of locomotive derailed) exhibits a stepwise increasing trend with financial losses. POSITON1 (car position in train first involved) and POSITON2 (causing car position in the train) from 125 to 135 are associated with the highest financial losses. These cars tend to be toward the rear of a typical Class I train (*49*).

By month, financial losses tend to peak in the summer and subside in the winter. In the U.S.A., grain harvesting and grain shipping by rail generally peak in the summer. Intuitively, peak demand leads to peak traffic with higher carloads, which increases the risk of accidents. T_loco shows that financial losses from accidents increases for

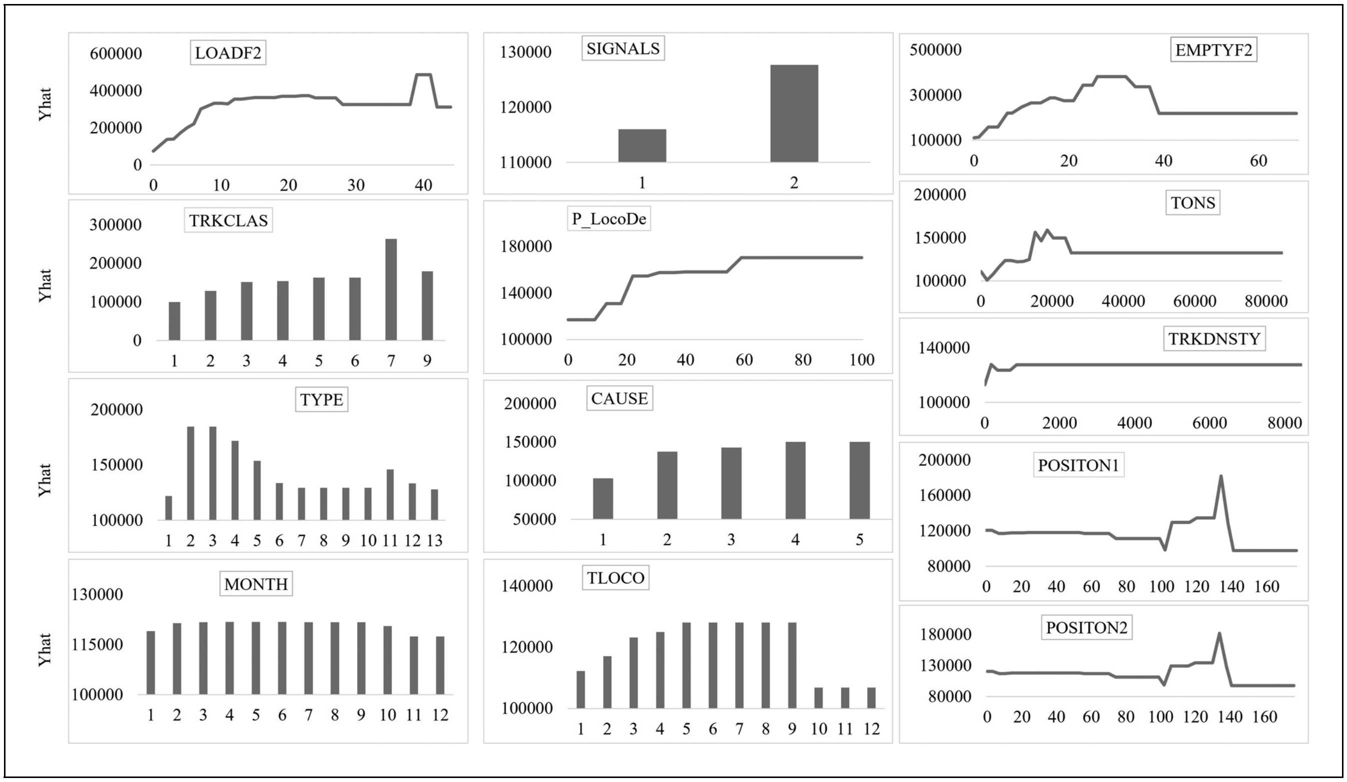

Figure 4. Partial-dependence plots of the predictor variables in the model.

trains that contained more than five locomotives. The variable weather shows that, compared to other weather conditions, snow is associated with a 0.03% increase in financial losses from accidents.

It is essential to highlight that because of the limitations in the knowledge provided by the data, the PDPs might not represent the true relationship between each variable and the predicted outcome. For instance, the financial losses from derailed empty freight cars are shown to be unchanged after 40 cars. Similarly, financial losses appear to be insensitive to TRKDNSTY. These problems could be the result of inconsistent data or missing data, which could be addressed by incorporating more data in the future.

Conclusion

The primary objective of this study was to determine the significant factors associated with Class I freight railroad financial losses from railway accidents and to rank the strength of those associations by using DM and ML techniques. Data between 2004 and 2018 from the REA database provided inputs for the analysis. To achieve the primary objective of the study, a comparative analysis of six ML algorithms determined the best model for the dataset. Tree-based ensemble models generally performed best. The XGBM proved to be the best model for analyzing railroad accident data that is highly imbalanced. The XGBM identified the significant factors associated with financial losses from railroad accidents. The results indicated that LOADF2 (number of derailed loaded freight cars), SIGNAL (type of territory signalization), and EMPTYF2 (number of derailed empty freight cars) were the top three factors with accuracy gains of 57%, 20%, and 10%, respectively, in predicting financial losses from railroad accidents. These results demonstrate the effectiveness of applying DM and ML techniques to high-volume and non-uniform data formats. The results suggest that railroads should prioritize safety investments that allow more trains to move freight on a signalized infrastructure.

Future work will explore and evaluate additional exogenous contributors to railroad accidents using a similar approach. The results will provide an opportunity to conduct a more comprehensive assessment of railroad accident contributors.

Author Contributions

The authors confirm their contribution to the paper as follows: study conception and design: N. Dhingra, R. Bridgelall, J.

Szmerekovsky; data collection: N. Dhingra; analysis and interpretation of results: N. Dhingra, R. Bridgelall, P. Lu, B. Bhardwaj; draft manuscript preparation: N. Dhingra, R. Bridgelall, P. Lu, B. Bhardwaj. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors express their deep gratitude to the following funding agencies' support: North Dakota State University and the Mountain-Plains Consortium (MPC), a University Transportation Center funded by the U.S. Department of Transportation under grant number DTRT13-G-UTC38.

ORCID iDs

Neeraj Dhingra https://orcid.org/0000-0001-9970-7185
Raj Bridgelall https://orcid.org/0000-0003-3743-6652
Pan Lu https://orcid.org/0000-0002-1640-3598
Joseph Szmerekovsky https://orcid.org/0000-0002-3355-9340
Bhavana Bhardwaj https://orcid.org/0000-0002-4379-1565

References

1. Association of American Railroads. Railroad 101- Freight Railroads Fact Sheet, 2020. https://www.aar.org/wp-content/uploads/2020/08/AAR-Railroad-101-Freight-Railroads-Fact-Sheet.pdf. Accessed May 23, 2021.
2. Federal Railroad Administration. *Monetary Threshold Notice*, 2021. https://railroads.dot.gov/forms-guides-publications/guides/monetary-threshold-notice. Accessed March 5, 2022.
3. Federal Railroad Administration. *Accident Data as Reported by Railroads*, 2018. https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/on_the_fly_download.aspx. Accessed March 7, 2019.
4. Wu, X., X. Zhu, G.-Q. Wu, and W. Ding. Data Mining With Big Data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, 2014, pp. 97–107.
5. Chung, Y.-S. Factor Complexity of Crash Occurrence: An Empirical Demonstration Using Boosted Regression Trees. *Accident Analysis & Prevention*, Vol. 61, 2013, pp. 107–118.
6. Chen, M., S. Mao, and Y. Liu. Big Data: A Survey. *Mobile Networks and Applications*, Vol. 19, 2014, pp. 171–209.
7. Li, L., S. Shrestha, and G. Hu. Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques. *Proc., 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, IEEE, London, 2017, pp. 363–370.
8. Abellán, J., G. López, and J. de Oña. Analysis of Traffic Accident Severity Using Decision Rules via Decision. *Expert Systems With Applications*, Vol. 40, No. 15, 2013, pp. 6047–6054.
9. Kohavi, R. Data Mining and Visualization. *Proc., 6th Annual Symposium on Frontiers of Engineering*, National Academy Press, Washington, D.C., 2001.
10. Barai, S. K. Data Mining Applications in Transportation Engineering. *Transport*, Vol. 18, No. 2, 2003, pp. 216–223.
11. Depaire, B., G. Wets, and K. Vanhoof. Traffic Accident Segmentation by Means of Latent Class Clustering. *Accident Analysis & Prevention*, Vol. 40, No. 4, 2008, pp. 1257–1266.
12. Sohn, S. Y., and S. H. Lee. Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea. *Safety Science*, Vol. 41, No. 1, 2003, pp. 1–14.
13. Ghomi, H., M. Bagheri, L. Fu, and L. F. Miranda-Moreno. Analyzing Injury Severity Factors at Highway Railway Grade Crossing Accidents Involving Vulnerable Road Users: A Comparative Study. *Traffic Injury Prevention*, Vol. 17, No. 8, 2016, pp. 833–841.
14. Ghomi, H., L. Fu, M. Bagheri, and L. F. Miranda-Moreno. Identifying Vehicle Driver Injury Severity Factors at Highway-Railway Grade Crossings Using Data Mining Algorithms. *Proc., 4th International Conference on Transportation Information and Safety (ICTIS)*, IEEE, Banff, AB, Canada, 2017, pp. 1054–1059.
15. JJ. *MAE and RMSE — Which Metric is Better?* 2016. https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. Accessed September 18, 2018.
16. Hu, S. R., C.-S. Li, and C.-K. Lee. Model Crash Frequency at Highway–Railroad Grade Crossings Using Negative Binomial Regression. *Journal of the Chinese Institute of Engineers*, Vol. 35, No. 7, 2012, pp. 841–852.
17. Kang, Y., and A. Khattak. Cluster-Based Approach to Analyzing Crash Injury Severity at Highway–Rail Grade Crossings. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2608: 58–69.
18. Lu, P., and D. Tolliver. Accident Prediction Model for Public Highway-Rail Grade Crossings. *Accident Analysis & Prevention*, Vol. 90, 2016, pp. 73–81.
19. Brown, D. E. Text Mining the Contributors to Rail Accidents. *IEEE Transactions on Intelligent Transportation Systems*. Vol. 17, No. 2, 2015, pp. 346–355.
20. Mirabadi, A., and S. Sharifian. Application of Association Rules in Iranian Railways (RAI) Accident Data Analysis. *Safety Science*, Vol. 48, No. 10, 2010, pp. 1427–1435.
21. Beshah, T., and S. Hill. Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. *Proc., AAAI Spring symposium series, AAAI Spring Symposium: Artificial Intelligence for Development*, Stanford, CA, 2010.

22. Mousa, S. R., P. R. Bakhit, O. A. Osman, and S. Ishak. A Comparative Analysis of Tree-Based Ensemble Methods for Detecting Imminent Lane Change Maneuvers in Connected Vehicle Environments. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 268–279.
23. Liu, X. Statistical Causal Analysis of Freight-Train Derailments in the United States. *Journal of Transportation Engineering, Part A: Systems*, Vol. 143, No. 2, 2017, p. 04016007.
24. Liu, X., M. R. Saat, X. Qin, and C. P. L. Barkan. Analysis of U.S. Freight-Train Derailment Severity Using Zero-Truncated Negative Binomial Regression and Quantile Regression. *Accident Analysis & Prevention*, Vol. 59, 2013, pp. 87–93.
25. Liu, X., M. R. Saat, and C. P. L. Barkan. Analysis of Causes of Major Train Derailment and their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*, 2012. 2289: 154–163.
26. Rahman, M. M., N. Haq, and R. M. Rahman. Machine Learning Facilitated Rice Prediction in Bangladesh. *Proc., Annual Global Online Conference on Information and Computer Technology*, IEEE, Louisville, KY, 2014, pp. 1–4.
27. Elsie Gyang, R., N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke, and N. J. Leeper. The Use of Machine Learning for the Identification of Peripheral Artery Disease and Future Mortality Risk. *Journal of Vascular Surgery*, Vol. 64, No. 5, 2016, pp. 1515–1522.
28. SAS. *Getting Started With SAS® Text Miner 12.1*, 2012. https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf. Accessed June 12, 2018.
29. García-Pedrajas, N., J. A. Romero Del Castillo, and G. Cerruela-García. A Proposal for Local k Values for k -Nearest Neighbor Rule. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 2, 2015, pp. 470–475.
30. Xie, Y., Y. Wang, A. Nallanathan, and L. Wang. An Improved K-Nearest-Neighbor Indoor Localization Method Based on Spearman Distance. *IEEE Signal Processing Letters*, Vol. 23, No. 3, 2016, pp. 351–355.
31. Breiman, L. Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32.
32. Liaw, A., and M. Wiener. Classification and Regression by RandomForest. *R News*, Vol. 2, No. 3, 2002, pp. 18–22.
33. Breiman, L. Bagging Predictors. *Machine Learning*, Vol. 24, No. 2, 1996, pp. 123–140.
34. Schapire, R. E., and Y. Singer. Improved Boosting Algorithms Using Confidence-Rated Predictions. *Proc., 11th Annual Conference on Computational Learning Theory, Association for Computing Machinery*, New York, NY, 1998. pp. 80–91.
35. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, Vol. 29, No. 5, 2001, pp. 1189–1232.
36. Friedman, J. H. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, Vol. 38, No. 4, 2002, pp. 367–378.
37. Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, and K. Chen. Xgboost: Extreme Gradient Boosting. *R Package Version 0.4-2*, Vol. 1, No. 4, 2015, pp. 1–4.
38. Bridgelall, R. *Lecture Notes: Introduction to Support Vector Machines*, 2017. Accessed May 26, 2018. https://assets.researchsquare.com/files/rs-1200362/v2_covered.pdf?c = 1643048728
39. Mountrakis, G., J. Lm, and C. Ogole. Support Vector Machines in Remote Sensing: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 66, No. 3, 2011, pp. 247–259.
40. Jahangiri, A., and H. A. Rakha. Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 5, 2015, pp. 2406–2417.
41. Yoonsuh, J., and J. Hu. A K-Fold Averaging Cross-Validation Procedure. *Journal of Nonparametric Statistics*, Vol. 27, No. 2, 2015, pp. 167–179.
42. Wong, T. T. Performance Evaluation of Classification Algorithms by k-Fold and Leave-One-Out Cross Validation. *Pattern Recognition*, Vol. 48, No. 9, 2015, pp. 2839–2846.
43. Liu, X., M. R. Saat, and C. P. L. Barkan. Freight-Train Derailment Rates for Railroad Safety and Risk Analysis. *Accident Analysis & Prevention*, Vol. 98, 2017, pp. 1–9.
44. Brownlee, J. *How to Configure k-Fold Cross-Validation*, 2020. https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/#:˜:text = The%20key%20configuration%20parameter%20for,evaluate%20models%20is%20k%3D10. Accessed January 18, 2022.
45. Guyon, I., and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1157–1182.
46. McCauliff, S. D., J. M. Jenkins, J. Catanzarite, C. J. Burke, J. L. Coughlin, J. D. Twicken, P. Tenenbaum, S. Seader, J. Li, and M. Cote. Automatic Classification of Kepler Planetary Transit Candidates. *The Astrophysical Journal*, Vol. 806, No. 1, 2015, p. 6.
47. Lundberg, S. *Toward Data Science: Interpretable Machine Learning With XGBoost*, 2018. https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27. Accessed December 16, 2018.
48. Greenwell, B. M. pdp: An R Package for Constructing Partial Dependence Plots, 2017. https://journal.r-project.org/archive/2017/RJ-2017-016/RJ-2017-016.pdf. Accessed May 18, 2018.
49. Cambridge Systematics, Inc. *National Rail Freight Infrastructure Capacity and Investment Study*, 2007. http://www.coaltrainfacts.org/docs/natl_freight_capacity_study.pdf. Accessed October 6, 2018.

The contents of this paper reflect the views of the authors, who are responsible for the facts and accuracy of the information presented.