Dean&Francis

ISSN 2959-6157

Prioritizing Recall: Recall-First Machine Learning for Traffic Accident Severity Detection under Class Imbalance

Geping Cai[1, *]

[1]College and Graduate School of Arts & Sciences, University of Virginia,The United States, 22903

*Corresponding author: caigeping9@gmail.com

Abstract:

Traffic safety agencies seek predictive models that identify high-risk crashes before traffic accidents occur, but real-world crash data are highly imbalanced, rendering the overall accuracy of machine learning models a misleading metric for predictive performance. This study investigates a recall-first approach to crash-risk classification in Virginia, arguing that missing a dangerous event carries a much higher cost than issuing an extra alert based on the data provided by Virginia Department of Transportation (VDOT). The modeling framed high-risk identification as a binary classification task and used standard tools, Logistic Regression and Random Forest, augmented with imbalance-aware training. Evaluation emphasized recall and precision, utilizing precision–recall curves and confusion matrices. Across experiments, recall-oriented training and thresholding consistently improved detection of high-risk cases relative to accuracy-optimized baselines, with expected trade-offs resulting in lower precision and overall accuracy. This paper further translates operational priorities into simple threshold policies, showing how agencies can tune models based on recall to align with resource constraints while minimizing costly misses. In conclusion, for safety-critical applications, recall should be the primary metric of success; furthermore, straightforward imbalance treatments—without complex model additions—can realign model performance with public-safety goals.

Keywords: Traffic Crash Risk Prediction; Class Imbalance; Recall-First Evaluation; Machine Learning Classification; Cost-Sensitive Decision Thresholds; Public-Safety Operations

ISSN 2959-6157

1. Introduction

Predictive models for traffic safety are increasingly used to identify high-risk crashes before they occur, leveraging roadway, temporal, and environmental context. Prior research identifies that machine learning can discern influential factors and predict crash hotspots, thereby motivating statewide, data-driven traffic safety prevention initiatives. However, within this body of literature, evaluation practices often downplay the significance of the minority class, which is the very event of greatest public-safety concern [1]. This paper focuses on recall-first crash-risk classification for Virginia, addressing pronounced class imbalance in the response variable data. The study explores how evaluation should prioritize recall over overall accuracy in safety-critical settings, and further examines how simple, widely used techniques (e.g., class weighting and post-training threshold selection) shape recall–precision trade-offs under class-imbalanced case. Concretely, this paper mainly addresses three questions in the study: (i) Why should evaluation frameworks prioritize recall for imbalanced data? (ii) How do class-imbalance-focused training and threshold choice shift operating points along the precision–recall curve and modify prediction model performance? (iii) How can agencies adjust thresholds according to cost ratios or alert-budget constraints?

Methodologically, this study employs two standard machine learning classification models, logistic regression and random forests, imbalance-focused training, and decision-threshold selection applied to statewide crash records with predictors, such as roadway, temporal, and environmental factors. Evaluation prioritized precision–recall analysis, macro/balanced metrics, and confusion matrices rather than accuracy—a metric that is often misleading, as it may reflect majority-class prevalence rather than true discriminative performance, ultimately failing to capture the underlying data pattern [2]. By centering recall and policy-aligned thresholds, this study provides practical guidance and hints for transportation agencies: how to tune alert volumes to match resources, how to minimize costly misses, and how to report metrics that align with public-safety objectives in real applications.

2. Data and Class Imbalance

2.1 Data Preparation and Preprocessing

The dataset utilized for model training and prediction in this study is provided by Virginia Department of Transportation (VDOT) and comprises 73 features with over 1 million records [3]. To prepare the dataset for predictive modeling, this study implemented a comprehensive data preprocessing pipeline designed to enhance model performance and reduce noise. The study first filtered the dataset to include only crashes from the years 2024 and 2025. Subsequently, irrelevant identifiers and features with consistently missing or non-informative values were removed, further defined the binary target variable High_Risk by mapping crash severity levels through encoding severity “0” as low risk (0) and all other severities as high risk (1) where high risk is identified as injury to people occurred in the accident. Rows with invalid or missing critical fields were removed to ensure data quality. For model input, 11 categorical variables pertaining to environmental and roadway conditions were selected, including weather, lighting, surface condition, road alignment, and work zone status. Each categorical feature was encoded using LabelEncoder, and all features were subsequently standardized using StandardScaler to normalize their ranges. Lastly, the filtered data was splitted into training, validation, and test sets using a $72\% - 8\% - 20\%$ split. Separate preprocessing pipelines for numerical features, employing median imputation and scaling, were applied. This preprocessing regimen ensured that all inputs were suitable for downstream machine learning models and upheld the integrity of the classification task.

2.2 Class Imbalance Across Response Variable

The dataset shows a moderate skew toward the non-high-risk class after reclassification with 68,992 instances labeled 0 (not high risk) versus 30,951 labeled 1 (high risk), roughly 7:3 overall (see Figure 1). This class imbalance implies that a simple classification model that always predicts the majority class would already attain about 70% accuracy without identifying any high-risk cases—a finding that further highlights why accuracy alone is misleading in such scenarios. Operationally, the 30% minority of high-risk class still represents a relatively large number of high-risk events. Consequently, the evaluation emphasizes recall and precision via PR analysis and utilizes thresholding calibrated to agency priorities, rather than maximizing overall accuracy.

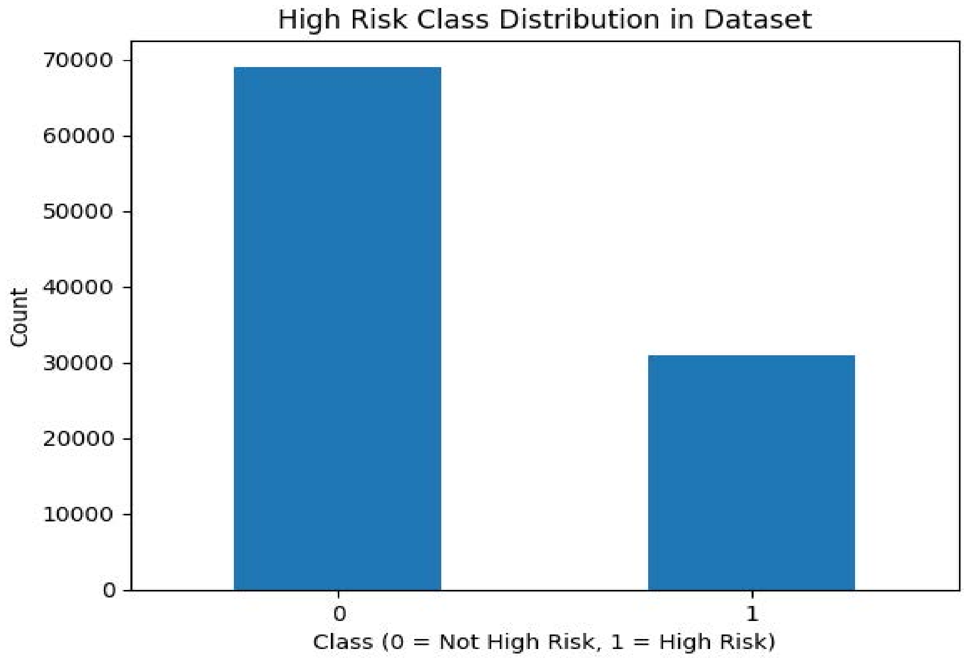

Fig. 1 High Risk Class Distribution in Dataset

3. Prioritizing Recall to Reduce False-Negative Harm

3.1 Policy and Real-World Consideration

The two types of errors examined in this paper carry distinctly different real-world consequences: a false negative (failing to predict a truly high-risk crash) risks severe injuries or fatalities with substantial societal costs, while a false positive (unnecessary alert provided by the model) primarily consumes staff time and modest operational resources. Highway safety practice in the U.S. already encodes this asymmetry: FHWA endorses severity-weighted evaluation such as the Equivalent Property Damage Only (EPDO) method, which assigns higher weights to fatal and injury crashes than to property damaged only events—a practice that strongly indicates missing severe events is far costlier than conducting extra screening and issuing alerts [4] [5]. Considering the Safe System Approach, the primary goal is to reduce deaths and serious injuries, thus, model selection and thresholding should emphasize recall (reducing false negatives) over raw accuracy that is usually used as an evaluation metric [6]. Furthermore, agencies in the real world always operate under finite resources and use data-driven and systematic strategies to prioritize risk across the network as the example of Rural Roadway by Federal Highway Administration—an approach that aligns with the core idea of selecting metrics and thresholds that control alert volumes while focusing on the most consequential events [7]. The stakes are important and cannot be ignored. NHTSA estimates the economic cost of U.S. motor-vehicle crashes at roughly $340 billion for 2019—a figure that underscores why minimizing missed severe events should guide model evaluation and deployment [8]. Given the safety-critical nature of traffic accident prediction and real-world consideration, the primary evaluation focus for the models was on recall scores. Failing to identify a high-risk crash (false negative) can have more severe real-world consequences than issuing a false warning (false positive). A high recall ensures that the models capture as many high-risk crashes as possible, even if it comes at the cost of increased false positives—a tradeoff that is acceptable in life-saving applications, where over-alerting is preferable to under-detecting dangerous scenarios.

3.2 Metric Choice: Recall-First Evaluation

In class-imbalanced and safety-critical settings, overall accuracy and even ROC/AUC might appear as a strong indicator of model performance measurement but actually covering the underlying truth of poor detection of the minority (high-risk) class, largely because the false positive (FP) rate is normalized by a large pool of negative instances. However, precision–recall (PR) analysis directly characterizes performance on the positive class: recall indicates the proportion of high-risk cases identified, while precision reflects the proportion of alerts that correspond to true high-risk cases. Prior work shows PR

ISSN 2959-6157

is more informative than ROC under class imbalance. Based on Saito & Rehmsmeier, the recall-first evaluation is chosen and reports Precision-Recall curve for Random Forest, as PR is more informative than ROC/AUC in imbalanced settings and better reflects the model's ability to "capture" the positive class [9]. Furthermore, according to Davis & Goadrich, Recall-First analysis with PR curves is preferred over ROC/AUC in this imbalanced setting, since ROC can appear strong despite poor positive class detection, while PR directly reflects high-risk case detection [10]. Accordingly, this study treats ROC/AUC and raw accuracy as less significant, focusing evaluation on recall and PR curves (PR only for Randon Forest).

4. Class Imbalance Handling

4.1 Class_weights in Logistic Regression and Random Forests

Class imbalance can be addressed during model training by assigning higher loss weights to the minority class, for example, via reweighted log-loss in Logistic Regression and class-weighted impurity/costs in Random Forests. This ensures errors on high-risk cases are penalized more heavily than errors on non-high-risk cases. Large empirical surveys show this family of cost-sensitive/weighted classifiers is competitive and often outperforms standard (unweighted) training on imbalanced datasets. Straightforward oversampling techniques (e.g., SMOTE, SMOTE+ENN) consistently yield better results than no imbalance treatment. This evidence supports reweighting the minority class as a simple, effective strategy for Logistic Regression (LR) and Random Forests (RF) to mitigate bias from class imbalance [11]. However, weighting is not always the optimal choice—merely one of the simplest. In tree ensembles specifically, researchers have proposed cost-sensitive or balanced-sampling variants of Random Forests, precisely because standard Random Forests tend to underdetect minority classes. In other words, class weights help but may not fully remove the bias without careful tuning [12].

5. Methods Implementation

5.1 Logistic Regression

A basic logistic regression model was implemented without hyperparameter tuning or class imbalance mitigation strategies to establish a performance baseline for the classification task. The model was trained using the default parameters and evaluated on the validation set. Although it achieved a validation accuracy of 68.8%, the model completely failed to identify any high-risk cases (class 1), exhibiting a recall and precision of 0.00 for that class. The confusion matrix revealed the model predicted only the majority class (not high risk), indicating its inability to handle class imbalance. This untuned logistic regression served as a useful starting point for error analysis, emphasizing the need for improved model design—particularly in addressing class imbalance.

To address the limitations of the baseline model, a tuned logistic regression model was implemented using grid search with 5-fold cross-validation. The best-performing hyperparameters were found to be C=0.01, class_weight='balanced', and solver='liblinear'—parameters that directly address severe class imbalance. This tuned model was evaluated on the test set, yielding more balanced performance. As shown in Table 1, although overall accuracy dropped to 48.8%, recall for the high-risk class (class 1) improved significantly to 0.63, with a corresponding precision of 0.33. This improvement is also reflected in the confusion matrix, which shows the model successfully identified 5,531 high-risk instances. These results demonstrate that model tuning and class weighting enhance the model's sensitivity to the minority class, confirming that addressing class imbalance is critical for risk-sensitive classification tasks.

Table 1. Logistic Regression Untuned vs. Tuned (Class 1 High Risk)

Model	Accuracy	Precision	Recall	TN	FP	FN	TP
Untuned	0.688	0.000	0.000	7,642	0	3,463	0
Tuned	0.488	0.334	0.633	8,022	11,007	3,203	5,531

5.2 Random Forests

A Random Forest model was also implemented with the same dataset to better capture non-linear patterns and feature interactions related to crashes. On the validation set, the model achieved an overall accuracy of 68.0%. However, while it demonstrated strong performance in identifying non-risky cases (class 0) with a recall of 0.97, it still struggled to detect high-risk crashes (class 1) without class imbalance handling, achieving a recall of only 0.04 for that class ("(recall = 0.97)" The confusion matrix further confirmed this issue, showing the model missed the vast majority of actual high-risk cases.

To fine tune the Random Forest model, this study experimented with class weighting, tree depth, and sample size constraints, conducting multiple trials to observe their impact on recall. One intermediate model using class_weight='balanced' and controlled tree complexity improved high-risk recall to 0.55 but lacked overall balance. To push recall even further, a final Random Forest model was developed using the following hyperparameters: n_estimators=1200, max_depth=20, min_samples_split=30, min_samples_leaf=20, max_features='sqrt', bootstrap=False, and a custom class_weight={0:1, 1:3}—a weight distribution determined by the class imbalance ratio in the dataset (see Table 2). These results further support the previous argument that addressing class imbalance is a key consideration for minority-class risk-event detection, as it significantly improves prediction model performance.

Table 2. Recall Score Progression During Random Forest Tuning

Trial #	Model Hyperparameter	Recall (High Risk)
1	Default Parameters (Baseline)	0.04
2	class_weight=`balanced`	0.55
3	class_weight={0:1, 1:2}	0.39
4	Final Tuned: class_weight={0:1, 1:2}	0.89
Final Tuned Model – Recall (Test Set)		0.89

For more information, the tuned Random Forest shows the classic precision–recall trade-off with the PR Curve as shown in Figure 2. At very strict thresholds, precision is near 0.9-1.0 on a small set of alerts—confirming the model ranks the riskiest cases highest. As the threshold lowers, recall rises smoothly toward 1 while precision declines, approaching the baseline at full recall. The curve remains above this baseline, indicating the model has discriminative power rather than relying on majority-class guessing. For a recall-first objective, the right half of the curve is particularly relevant, as it offers an acceptable trade-off when the cost of missed high-risk crashes outweighs the cost of extra alerts.

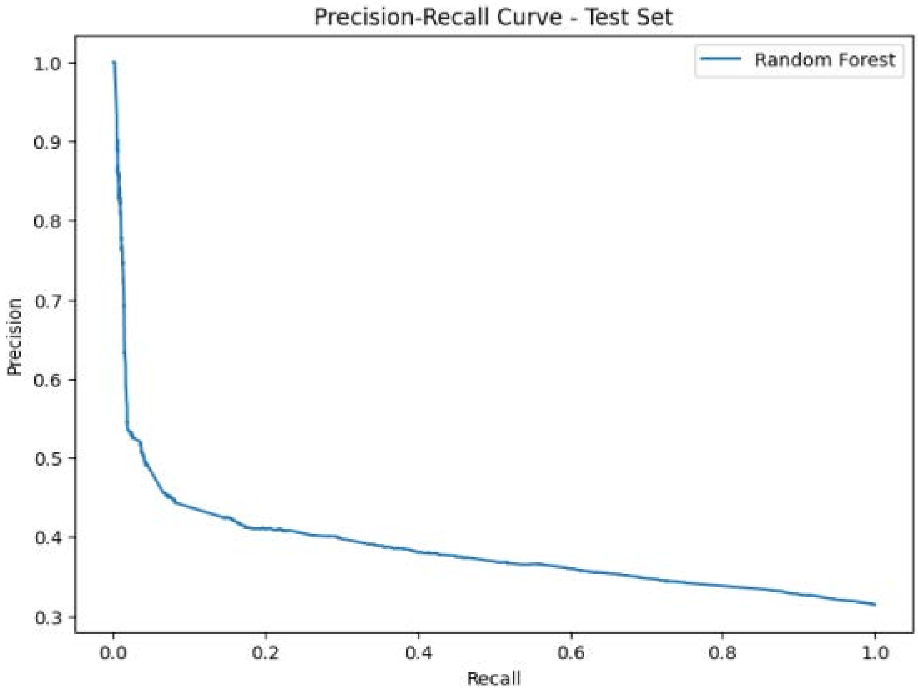

Fig. 2 Tuned Random Forest PR Curve

ISSN 2959-6157

6. Conclusion

This study frames traffic crash-severity prediction in Virginia as a recall-first problem on imbalanced data. A baseline logistic regression missed high-risk cases, but the model's ability to detect minority-class cases was significantly enhanced through the integration of class weights and post-training threshold adjustment. Moreover, a tuned class-weighted Random Forest delivered strong performance characterized by high recall at an acceptable level of precision—a result that better aligns with FHWA's cost asymmetry, where false negatives outweigh false positives. Focusing evaluation on Recall and PR Curve provides a clearer view of safety impact and supports threshold tuning aligned with operational capacity. The approach is simple and actionable for agencies, and future work will explore the finer balance between recall and precision, as well as address the study's limitations.

References

[1] Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. Computers, 10(12), 157. https://doi.org/10.3390/computers10120157

[2] Ghanem, M., Ghaith, A. K., El-Hajj, V. G., Bhandarkar, A., de Giorgio, A., Elmi-Terander, A., & Bydon, M. (2023). Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. Brain sciences, 13(12), 1723. https://doi.org/10.3390/brainsci13121723

[3] Virginia Department of Transportation. Full Crash Dataset. https://www.virginiaroads.org/datasets/VDOT::full-crash-1/explore?layer=0&location=0.003769%2C-79.499811%2C0.00&showTable=true.

[4] FHWA Safety Program. (n.d.-a). Crash costs for highway safety analysis. U.S. Department of Transportation Federal Highway Administration. https://safety.fhwa.dot.gov/hsip/docs/fhwasa17071.pdf

[5] U.S. Department of Transportation. (2004, August). Signalized intersections: Informational guide. FHWA. https://www.fhwa.dot.gov/publications/research/safety/04091/

[6] What is a safe system approach?. U.S. Department of Transportation. (2025, January 14). https://www.transportation.gov/safe-system-approach

[7] EDC-5 Innovations (2019-2020). EDC-5 Innovations | Federal Highway Administration. (2023, May 25). https://www.fhwa.dot.gov/innovation/everydaycounts/edc_5/

[8] Blincoe, L., Miller, T., Wang, J.-S., Swedler, D., Coughlin, T., Lawrence, B., Guo, F., Klauer, S., & Dingus, T. (2023, February). The economic and societal impact of motor vehicle crashes, 2019 (Revised) (Report No. DOT HS 813 403). National Highway Traffic Safety Administration.

[9] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.

[10] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

[11] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250, 113-141.

[12] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24.