CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21900

Operational Evaluation of Rare-Event Prediction in Socio-Economic Data: Prevalence Regimes, Budget-Constrained Targeting, and Distribution Shift

Jaime Villanueva-García[1], Luis Javier García Villalba[1], Ignacio Moral-Arce[2]

[1]Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Complutense University of Madrid (UCM), Spain, [2]Institute of Fiscal Studies (IEF), Ministry of Finance, Madrid, Spain

How to cite: Villanueva-García, J.; García Villalba, L. J.; Moral-Arce, I. (2026). Operational Evaluation of Rare-Event Prediction in Socio-Economic Data: Prevalence Regimes, Budget-Constrained Targeting, and Distribution Shift. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.21900

Abstract

Artificial intelligence increasingly supports decision-making in socio-economic contexts characterized by rare but consequential events. However, model evaluation often relies on global discrimination metrics that may not reflect operational realities where interventions are constrained by limited budgets or administrative capacity.

This paper proposes a deployment-oriented evaluation framework for rare-event prediction based on upper-tail performance under fixed intervention rates. Using socio-economic, financial, and administrative datasets spanning moderate- and low-prevalence regimes, we compare logistic regression, random forests, and gradient boosting models with and without isotonic calibration.

Results show that the interaction between event prevalence and intervention budgets strongly shapes model ranking behavior. Under low prevalence, small intervention fractions sharply differentiate models in the extreme upper tail, whereas under moderate prevalence performance differences remain compressed. In this benchmark, isotonic calibration does not consistently improve budget-based outcomes, and a simple feature-based out-of-distribution stress test can substantially degrade upper-tail recall.

These findings highlight the importance of regime-aware evaluation when deploying AI systems in socio-economic policy and administrative contexts.

Keywords: *Rare-event prediction; Imbalanced learning; Budget-constrained decision-making; Model evaluation; Distribution shift; Applied artificial intelligence*

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

1. Introduction

Class imbalance is a pervasive challenge in applied machine learning, particularly when the positive class corresponds to rare but consequential events such as fraud detection, credit default, eligibility for public programs, or administrative transitions like retirement. In these settings, predictive models are often used to support socio-economic decision-making processes where identifying a small number of high-risk cases is operationally important.

A substantial literature addresses the algorithmic challenges posed by imbalanced data. Early work emphasized resampling methods, cost-sensitive learning, and ensemble techniques to mitigate skewed class distributions (He & Garcia, 2009; Krawczyk, 2016; Haixiang et al., 2017). More recent studies continue to focus on improving discrimination metrics such as ROC-AUC, PR-AUC, and F-measures through increasingly sophisticated models (Chen et al., 2024; Wainer, 2024; Jiang et al., 2023).

However, many real-world deployments operate under strict operational constraints. Interventions are often limited to a fixed fraction of the population due to budgetary or administrative restrictions. In such settings, the relevant question is not only how well a model discriminates globally, but how effectively it ranks individuals in the extreme upper tail of predicted risk. Evaluation therefore becomes conditional on a fixed intervention rate.

While prior research has largely examined imbalanced classification from an algorithmic perspective (e.g., Galar et al., 2012; Mushava & Murray, 2022), comparatively less attention has been devoted to how evaluation behavior changes across prevalence regimes and operational constraints. In particular, the interaction between event prevalence (π), intervention budget (k), and upper-tail ranking efficiency remains insufficiently explored. This issue is especially relevant in rare-event contexts with prevalence below 5%, where global discrimination metrics may fail to capture deployment-relevant performance.

Building on previous benchmarking work in imbalanced classification (Villanueva-García et al., 2026), this study shifts the focus from algorithmic comparison to deployment-oriented evaluation. Rather than proposing new predictive models, we analyze how model ranking behavior and calibration performance interact with prevalence regimes and operational budgets. Our empirical analysis covers three contexts: (i) moderate imbalance in socio-economic datasets, (ii) moderate imbalance in financial default prediction, and (iii) a low-prevalence administrative dataset with event rates close to 2.3%.

Our central claim is that model evaluation is regime-dependent. As prevalence decreases, upper-tail ranking becomes more informative for operational decision-making, and metrics such as Recall@k provide stronger model differentiation. Calibration and distributional perturbations may also interact with prevalence in ways that affect both probability quality and deployment-relevant recall.

Article in Press

Contributions. This paper makes four contributions to the evaluation of rare-event prediction models under operational constraints:

- First, we formalize a deployment-oriented evaluation framework based on fixed intervention budgets and upper-tail metrics such as Recall@k.
- Second, we show empirically that the informativeness of upper-tail evaluation depends on the interaction between prevalence and intervention budgets.
- Third, using logistic regression, random forests, and gradient boosting models with and without isotonic calibration, we show that isotonic calibration does not consistently improve budget-based outcomes in the present benchmark.
- Finally, we demonstrate that a simple feature-based out-of-distribution stress test can substantially degrade operational performance, highlighting the need to monitor deployment-relevant metrics beyond global discrimination measures.

2. Data and Empirical Design

To examine how evaluation behavior varies across prevalence regimes and deployment settings, we analyze three empirical contexts spanning moderate- and low-prevalence conditions.

First, we use the Adult income dataset (OpenML/UCI), where the target variable indicates whether annual income exceeds $50,000. This dataset represents a moderate imbalance regime with a positive-class prevalence of approximately 24%. Second, we analyze the Taiwanese credit default dataset, where the event of interest is default in the following month and the prevalence is around 22%. Third, we consider an administrative dataset derived from social security records, where the target variable indicates retirement transitions. This setting represents a low-prevalence context with event rates close to 2.3%.

To examine subpopulation heterogeneity, the administrative dataset is split into male and female cohorts with similar prevalence but different covariate distributions, allowing us to assess differences in upper-tail targeting performance across groups.

All datasets are evaluated using a consistent modeling pipeline. Models are trained with a stratified 75–25 train–test split to preserve class prevalence. We compare three widely used classifier families: logistic regression, random forests, and gradient boosting (XGBoost). For each model, we evaluate both raw predicted probabilities and a post-hoc isotonic calibration variant.

Preprocessing includes median imputation for numerical features, most-frequent imputation for categorical variables, one-hot encoding, and feature scaling. No dataset-specific hyperparameter optimization is performed, allowing evaluation differences to reflect regime

Article in Press

effects rather than algorithm tuning. For calibrated models, isotonic calibration is fitted through internal cross-validation within the training data.

Evaluation combines global and deployment-oriented metrics. Global discrimination is primarily assessed using PR-AUC and ROC-AUC, while probability quality is assessed via the Brier score. Deployment performance is evaluated using Recall@k and Precision@k for $k \in \{1\%, 5\%, 10\%\}$. Recall@k measures event coverage within the intervention budget, while Precision@k measures targeting purity among selected cases. Under random targeting, expected Recall@k equals k.

Finally, we introduce a simple feature-based out-of-distribution stress test. For each dataset, observations above the 80th percentile of a selected numerical feature are held out as an OOD test set while models are trained on the remaining data. In the Adult and Credit datasets this feature corresponds to age, while in the administrative dataset it is selected among available continuous variables. This perturbation provides a controlled way to evaluate how upper-tail recall and discrimination respond to shifts in covariate support.

3. Results

This section evaluates how model performance varies across prevalence regimes, intervention budgets, calibration choices, and feature-based distribution shifts. We focus on deployment-oriented metrics, particularly Recall@k and Precision@k for $k \in \{1\%, 5\%, 10\%\}$, and examine how model comparisons change in the operational region of the score distribution rather than through global discrimination metrics alone.

3.1. Regime Interaction: Upper-Tail Ranking under Moderate and Low Prevalence

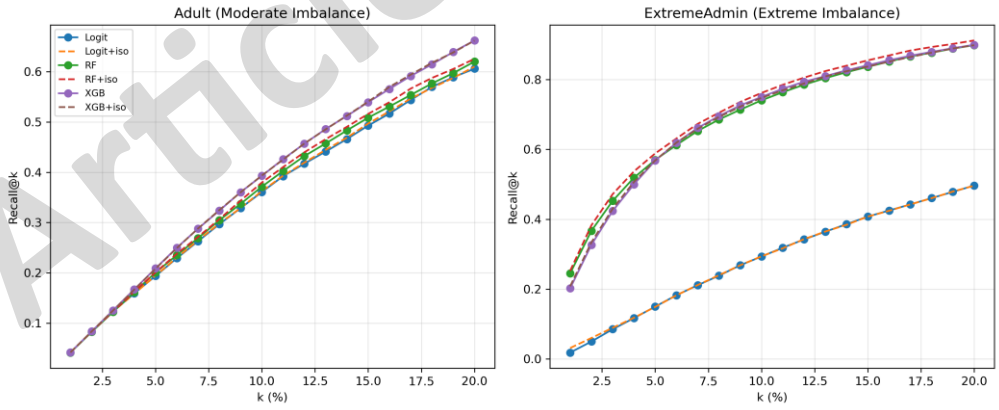


Figure 1. Recall@k curves under moderate- and low-prevalence regimes. Source: Authors.

Article in Press

Figure 1 reports Recall@k curves for $k = 1\%$ to 20% in two representative settings: the Adult dataset, which reflects moderate imbalance, and the ExtremeAdmin_Male dataset, which represents a low-prevalence administrative context with event rates close to 2.3%.

In the Adult dataset, Recall@1% remains structurally compressed across models, around 3–4%, and meaningful differences emerge only at larger intervention budgets ($k \geq 5\%$). The best-performing configuration (XGB + isotonic) achieves Recall@1% = 0.042 and Recall@10% = 0.394.

By contrast, upper-tail recall becomes substantially more discriminative in the low-prevalence administrative dataset. In the male cohort, calibrated random forests reach Recall@1% = 0.252 and Recall@10% = 0.761. Similar patterns appear in the female cohort, where Recall@1% = 0.206 and Recall@10% = 0.750.

These results indicate that evaluation behavior depends strongly on the underlying prevalence regime. In low-prevalence contexts, even small intervention budgets produce substantial separation between models in the extreme upper tail, whereas in moderate-prevalence datasets such as Adult and Credit Default upper-tail recall remains more compressed at small values of k.

Table 1 summarizes selected tree-based model configurations across datasets. While ROC-AUC remains relatively high across contexts, upper-tail recall varies considerably. In administrative datasets Recall@1% exceeds 20%, whereas in moderate-prevalence datasets it remains below 5%, illustrating that global discrimination metrics may fail to capture operational model value when interventions target only a small fraction of the population.

Table 1. Selected tree-based model configurations across datasets. Source: Authors.

Dataset	Model	PR_AUC	Recall@1%	Precision@1%	Recall@10%
Adult	XGB + isotonic	0.834	0.042	1.000	0.394
Credit	XGB	0.554	0.037	0.813	0.312
Extreme_Male	RF + isotonic	0.394	0.252	0.593	0.761
Extreme_Female	RF + isotonic	0.324	0.206	0.479	0.750

3.2. Cohort Heterogeneity in Upper-Tail Targeting

To examine whether upper-tail performance varies across subpopulations beyond base-rate differences, we compare model outcomes across administratively separated cohorts within the low-prevalence setting. Figure 2 reports Recall@1% for male and female populations using calibrated tree-based models.

Article in Press

Despite similar prevalence across cohorts, upper-tail recall differs systematically. For calibrated random forests, Recall@1% equals 0.252 in the male cohort and 0.206 in the female cohort. A similar pattern appears for gradient boosting models, with Recall@1% values of 0.208 and 0.177 respectively.

These results indicate that differences in covariate structure, rather than prevalence alone, can influence ranking efficiency in the extreme upper tail. Consequently, operational targeting performance may vary across subpopulations, suggesting that deployment-oriented evaluation should consider cohort-level variation rather than relying exclusively on aggregate metrics.

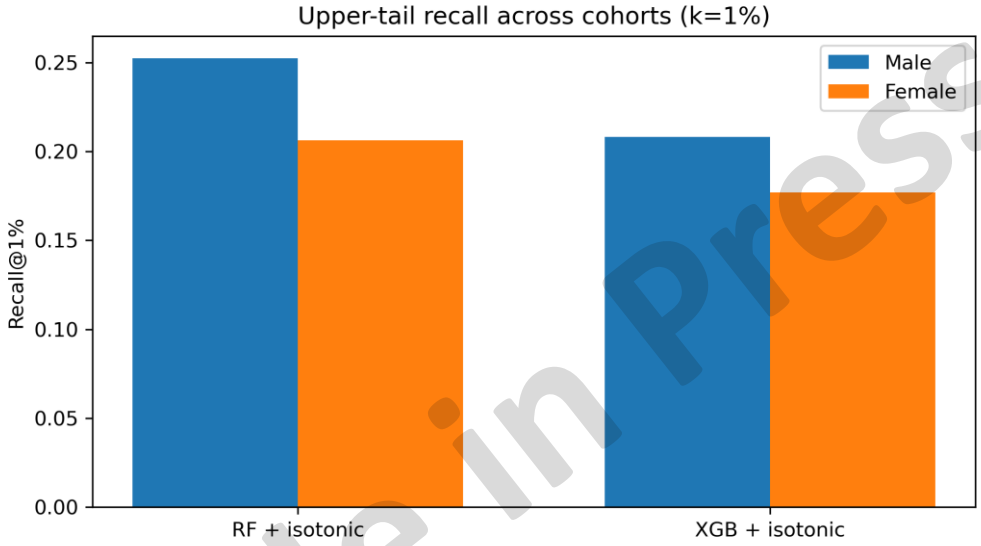


Figure 2. Recall@1% across male and female administrative cohorts. Source: Authors.

3.3. Calibration and Sensitivity under Feature-Based Distribution Shift

We next evaluate model performance under a simple feature-based distribution shift. For each dataset, observations above the 80th percentile of a selected numerical feature are held out as an out-of-distribution (OOD) test set while models are trained on the remaining observations. This design provides a controlled perturbation of covariate support rather than simulating realistic temporal drift. Table 2 compares in-distribution (random split) and OOD performance for selected models. Across datasets, distributional shift reduces both discrimination metrics and upper-tail recall. In the credit default dataset, calibrated gradient boosting shows a substantial decline under OOD conditions, with Recall@10% dropping from 0.309 to 0.163 and PR-AUC from 0.553 to 0.329.

Article in Press

The effect is even stronger in the low-prevalence administrative dataset. In the male cohort, calibrated random forests achieve Recall@10% = 0.761 in the random split but only 0.251 under OOD conditions. Similar declines appear across models and cohorts, indicating that upper-tail targeting performance can deteriorate sharply when covariate distributions change.

Overall, these results highlight the sensitivity of deployment-oriented evaluation to shifts in covariate support. Strong in-distribution upper-tail performance may therefore not transfer to shifted environments, emphasizing the importance of stress testing and monitoring deployment-relevant metrics in operational settings.

Table 2. In-distribution versus feature-based out-of-distribution performance. Source: Authors.

Dataset	Model	Recall@10% (Random)	Recall@10% (OOD)	PR_AUC (Random)	PR_AUC (OOD)
Adult	RF + isotonic	0.380	0.295	0.789	0.773
Adult	XGB + isotonic	0.394	0.311	0.834	0.818
Credit	RF + isotonic	0.311	0.273	0.548	0.540
Credit	XGB + isotonic	0.309	0.163	0.553	0.329
Extreme_Male	RF + isotonic	0.761	0.251	0.394	0.118
Extreme_Male	XGB + isotonic	0.747	0.242	0.328	0.135
Extreme_Female	RF + isotonic	0.750	0.180	0.324	0.119
Extreme_Female	XGB + isotonic	0.693	0.268	0.264	0.181

4. Discussion and Implications

This study shows that rare-event model evaluation is strongly influenced by the interaction between prevalence regimes and operational deployment constraints. When interventions target only a small fraction of the population, upper-tail metrics such as Recall@k provide a more decision-relevant assessment than global discrimination measures. Our results indicate that the relationship between prevalence (π) and intervention budget (k) plays a central role in shaping operational evaluation behavior. Under moderate prevalence, small targeting budgets produce compressed performance differences, whereas in low-prevalence settings the extreme upper tail becomes highly discriminative and model choice can materially affect outcomes.

Article in Press

The findings also show that improvements in probability calibration do not automatically translate into better deployment performance. Post-hoc isotonic calibration may improve probability quality in-distribution, yet it does not consistently increase upper-tail recall and may coincide with substantial performance degradation under feature-based distribution shifts. Calibration strategies should therefore be evaluated within the operational decision region where models are applied, rather than relying solely on aggregate probability-quality or discrimination metrics. From a policy perspective, predictive models should be assessed under the same decision regime in which they are deployed. In many socio-economic applications, models prioritize cases for intervention, auditing, or resource allocation, making upper-tail evaluation a more realistic measure of decision-support value. Future research may extend this framework to temporal shifts, alternative calibration approaches, and cost-sensitive utility formulations tailored to specific policy contexts.

References

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10759-6

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics: Part C (Applications and Reviews)*, 42(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2160949

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Jiang, C., Lu, W., Wang, Z., & Ding, Y. (2023). Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. *Expert Systems with Applications*, 213, 118878. https://doi.org/10.1016/j.eswa.2022.118878

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Mushava, J., & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, 202, 117233. https://doi.org/10.1016/j.eswa.2022.117233

Villanueva-García, J., Moral-Arce, I., & García Villalba, L. J. (2026). A cumulative, pipeline-oriented framework for extreme imbalanced classification: Large-scale benchmarking of 4,100 rare-event models. *Expert Systems with Applications, 312*, 131366. https://doi.org/10.1016/j.eswa.2026.131366

Article in Press

Wainer, J. (2024). An empirical evaluation of imbalanced data strategies from a practitioner's point of view. *Expert Systems with Applications*, 256, 124863. https://doi.org/10.1016/j.eswa.2024.124863

Article in Press