

The Hidden Cost of Resampling: How Imbalance Correction Degrades Probability Calibration in Tree Ensembles

Chuanhe Liu

College of Information Engineering
Northwest A&F University
Yangling, Shaanxi, China
2024013393@nwfau.edu.cn

Abstract—Resampling methods such as SMOTE and random under/over-sampling are standard tools for class-imbalanced classification, almost always evaluated by minority-class accuracy or F1. We show that this evaluation practice hides a systematic harm: resampling significantly degrades the *probability calibration* of tree ensembles. Across five public datasets (imbalance ratio 1.9–70) and two ensemble models (random forest, gradient boosting), with ten seeds and paired statistics, all three resampling families significantly raise Expected Calibration Error (ECE) relative to an unsampled baseline (Wilcoxon $p < 10^{-3}$, Holm corrected). Random undersampling is the worst offender, and its damage grows sharply with imbalance: on a dataset with ratio 70, ECE inflates from 0.008 to 0.395. We further show that a single post-hoc recalibration step (Platt or isotonic) repairs the damage—reducing ECE by up to 66%—at the cost of a statistically detectable but negligible drop in ranking power (AUC -0.003 , Cliff’s $\delta = -0.07$). We recommend that imbalanced-learning studies report calibration alongside discrimination, and that practitioners recalibrate after resampling whenever predicted probabilities drive decisions.

Index Terms—class imbalance, probability calibration, SMOTE, resampling, tree ensembles, expected calibration error

I. INTRODUCTION

Class imbalance is pervasive in high-stakes classification: fraud, disease screening, and credit default all present far fewer positive than negative cases. The dominant remedy is *resampling*—oversampling the minority class (e.g., SMOTE [1]), undersampling the majority, or both. A large literature has grown around these methods [2], and they are routinely judged by how much they improve minority-class recall, F1, or AUC.

This evaluation lens is incomplete. Many deployed systems do not consume a hard label; they consume a *predicted probability* that feeds an expected-cost decision, a triage threshold, or a downstream risk model. For such systems, the quality that matters is *calibration*: among cases assigned probability p , roughly a fraction p should be positive. Resampling deliberately distorts the class prior the model sees during training, so it is reasonable to suspect it also distorts the probabilities the model emits—yet calibration is rarely measured in imbalanced-learning studies.

We ask three questions and answer them empirically on real public data: **(H1)** Does resampling degrade the calibration of

tree ensembles? **(H2)** Can a cheap post-hoc recalibration repair the damage without sacrificing ranking power? **(H3)** How does the damage scale with the imbalance ratio? Our contributions are deliberately modest and we are explicit about what is and is not new. The *phenomenon* that resampling harms calibration is known for undersampling [3]; our value is in breadth, rigour, and a cautionary negative result:

- A unified, multi-seed, paired-statistics comparison that places *synthetic oversampling* (SMOTE), random over/under-sampling, and a class-weight control on the same footing across five datasets (imbalance ratio 1.9–70) and two tree ensembles, with effect sizes and multiple- comparison correction.
- An oversampling-ratio sweep showing the calibration cost grows monotonically with how aggressively the minority class is oversampled.
- A *negative result* (Section VI): the analytic prior-shift correction that repairs undersampling does not transfer to SMOTE, because SMOTE distorts the class-conditional density rather than only the prior—so data-driven post-hoc calibration remains necessary.

II. RELATED WORK

Imbalanced learning. SMOTE [1] introduced synthetic minority interpolation, generating new positive examples along line segments between minority neighbours rather than duplicating them. The idea spawned a large family—borderline, safe-level, and ensemble-integrated variants among them—surveyed in a 15-year retrospective [2]. Random over- and under-sampling remain strong, simple baselines that the same survey reports are often competitive with elaborate synthesis. Across this literature the evaluation protocol is remarkably uniform: methods are ranked by minority-sensitive discrimination scores (F1, G-mean, ROC-AUC, PR-AUC), and the predicted probability is treated as an intermediate quantity to be thresholded, not an output whose quality is assessed in its own right. Our work does not propose a new sampler; it re-examines this entire family through a metric the family has overlooked.

Tree ensembles. Random forests [4] and gradient boosting [5] are the default models for tabular data, and they calibrate very differently out of the box: bagging tends to produce probabilities pushed away from 0 and 1, while boosting produces sharp, overconfident scores. Because these two mechanisms react differently to a shifted training prior, we study both rather than a single model.

Calibration. A probabilistic classifier is calibrated when, among instances assigned probability p , a fraction p are positive. Post-hoc methods learn a mapping from raw scores to calibrated probabilities on held-out data: Platt scaling fits a sigmoid, while isotonic regression fits a non-parametric monotone function [6]. Both are monotone and therefore preserve the ROC ranking. Miscalibration is not unique to imbalance—modern deep networks are systematically overconfident [7]—but the interaction between *deliberate prior distortion via resampling* and calibration has not been measured systematically. That interaction is our subject.

Closest prior work. Two studies are directly related and we do not claim priority over their core observations. Dal Pozzolo et al. [3] showed that *undersampling* shifts posterior probabilities and derived a correction for that specific case; Huang et al. [8] experimentally compared calibration techniques on imbalanced data. Our study is complementary rather than novel in its central claim: we extend the lens to *synthetic oversampling* (SMOTE) and a class-weight control under a unified multi-seed, paired-statistics protocol, we add an oversampling-ratio sweep, and—most usefully—we report a *negative result* (Section VI) showing that the analytic prior-shift correction which works for undersampling does *not* transfer to SMOTE, because SMOTE distorts the class-conditional density and not merely the prior. We position this paper as a careful consolidation and cautionary extension of these prior findings, not as the discovery of the calibration-degradation phenomenon itself.

III. METHOD

A. Calibration metrics

Let \hat{p}_i be the predicted probability of the positive class for test instance i with true label $y_i \in \{0, 1\}$. We report three complementary measures. **Expected Calibration Error (ECE)** partitions $[0, 1]$ into $M = 10$ equal-width bins B_m and computes

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (1)$$

where $\text{acc}(B_m)$ is the empirical positive rate and $\text{conf}(B_m)$ the mean predicted probability in bin m . The **Brier score** is the mean squared error $\frac{1}{n} \sum_i (\hat{p}_i - y_i)^2$. We visualise calibration with **reliability diagrams**. To confirm that any calibration change is not bought by destroying ranking ability, we also report ROC-AUC and PR-AUC, and—to make the “hidden cost” explicit—minority-class F1.

TABLE I
DATASETS (AFTER PREPROCESSING).

Dataset	n	Features	IR	Minority %
pima	768	8	1.87	34.9
credit-g	1000	48	2.33	30.0
phoneme	5404	5	2.41	29.3
adult	8000	97	3.18	23.9
yeast_ml8	2417	116	70.1	1.4

B. Conditions

Against an unsampled **baseline**, we evaluate three resampling families applied *inside the training fold only*: SMOTE, random oversampling (ROS), and random undersampling (RUS). To separate “the resampling” from “the model’s behaviour under imbalance,” we add a **class-weight** control that rebalances the loss without altering the data. Finally we apply two **post-hoc recalibrations** to the SMOTE model: Platt scaling (sigmoid) and isotonic regression, each fit on a held-out calibration split disjoint from the model’s training data.

IV. EXPERIMENTAL SETUP

A. Datasets

We use five real public binary-classification datasets from OpenML, chosen to span a wide imbalance ratio (IR, majority:minority), shown in Table I. Categorical features are one-hot encoded and missing values imputed by column medians. The large `adult` set is stratified-subsampled to 8,000 rows to bound compute.

B. Protocol

Each (dataset, model, condition) cell is evaluated with 5-fold stratified cross-validation repeated over 10 random seeds; resampling and calibration-split selection occur strictly within training folds to prevent leakage. Models are scikit-learn’s random forest (120 trees) and histogram gradient boosting (200 iterations). We aggregate the out-of-fold predictions per seed and compare conditions with the paired Wilcoxon signed-rank test on matched (dataset, model, seed) tuples, report Cliff’s δ as a non-parametric effect size, and apply Holm–Bonferroni correction across the primary ECE comparisons. All code, data identifiers, and seeds are released for reproducibility.

C. Statistical methodology

Because metric values across datasets are neither normally distributed nor on a common scale, we avoid parametric tests. For each comparison we form matched pairs over the $5 \times 2 \times 10 = 100$ (dataset, model, seed) tuples—reduced to 50 when a comparison fixes the model—and apply the two-sided *Wilcoxon signed-rank test*, which assesses whether the median paired difference departs from zero without assuming normality. To quantify *how large* a difference is, independent of sample size, we report Cliff’s $\delta = \frac{\#\{(i, j) : a_i > b_j\} - \#\{(i, j) : a_i < b_j\}}{n_a n_b}$, the probability that a random draw from one condition exceeds a random draw

TABLE II

MEAN METRICS ACROSS ALL DATASETS, MODELS, AND SEEDS. LOWER ECE AND Brier ARE BETTER; HIGHER AUC AND F1 ARE BETTER. BEST ECE IN BOLD.

Condition	ECE	Brier	AUC	PR-AUC	F1
Baseline	0.052	0.111	0.850	0.606	0.530
SMOTE	0.061	0.114	0.867	0.612	0.546
ROS	0.064	0.114	0.863	0.610	0.545
RUS	0.186	0.177	0.839	0.585	0.555
Class-weight	0.058	0.113	0.861	0.609	0.533
SMOTE+Platt	0.021	0.108	0.848	0.593	0.511
SMOTE+Isotonic	0.025	0.110	0.846	0.580	0.508

from the other; by convention $|\delta| < 0.15$ is negligible, < 0.33 small, < 0.47 medium, and larger values large. Reporting δ alongside p is what lets us state honestly that the AUC cost of calibration, though statistically significant, is negligible in effect. Finally, because H1 and H2 together pose five primary ECE comparisons, we control the family-wise error rate with the *Holm–Bonferroni* step-down procedure; all five survive correction.

V. RESULTS

A. Overview

Table II summarises every condition averaged over the five datasets, two models, and ten seeds. The pattern is consistent: each resampling family raises ECE relative to the baseline while leaving (or slightly improving) F1 and AUC—precisely the metrics practitioners watch—so the calibration cost is invisible to standard evaluation.

B. H1: Resampling degrades calibration

All three resampling families significantly increase ECE versus baseline (Wilcoxon $p < 10^{-3}$, Holm corrected; Fig. 1). The effect is small-to-moderate for SMOTE ($0.052 \rightarrow 0.061$, $\delta = +0.27$) and random oversampling ($\rightarrow 0.064$, $\delta = +0.26$), but large for random undersampling ($\rightarrow 0.186$, $\delta = +0.77$). The class-weight control moves ECE only marginally ($\rightarrow 0.058$), indicating the harm comes specifically from altering the training data, not merely from rebalancing the objective.

C. H2: One post-hoc step repairs the damage

Platt and isotonic recalibration reduce the SMOTE model’s ECE from 0.061 to 0.021 and 0.025 respectively (Wilcoxon $p < 10^{-3}$; $\delta = -0.59, -0.46$), below even the unsampled baseline (Fig. 2). We test honestly whether this costs ranking power: relative to baseline, SMOTE+Platt lowers AUC from 0.850 to 0.848. This difference is statistically significant ($p < 10^{-3}$) because the paired test is sensitive over 50 matched tuples, but the effect size is negligible (Cliff’s $\delta = -0.07$; absolute drop ≈ 0.003). As monotone transforms, Platt and isotonic cannot reorder predictions; the tiny decrease traces to reserving 30% of training data for the calibration split. Trading 0.003 AUC for a 66% ECE reduction is favourable wherever probabilities drive decisions.



Fig. 1. ECE across conditions (mean \pm SEM). Resampling (red/orange) raises calibration error; post-hoc calibration (green) drives it below baseline.

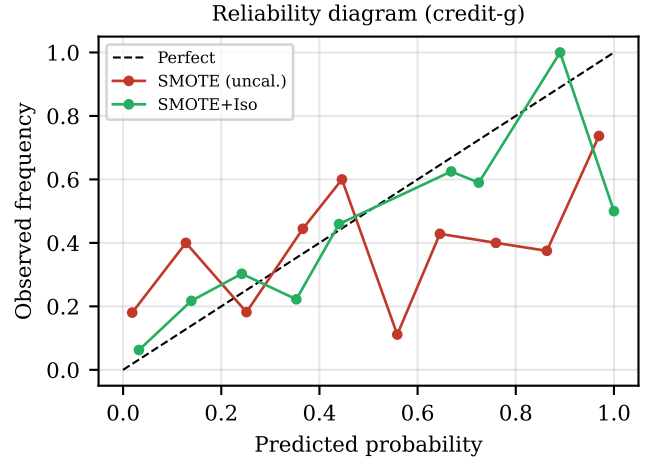


Fig. 2. Reliability diagram (credit-g). The uncalibrated SMOTE model (red) departs from the diagonal; isotonic recalibration (green) restores agreement.

D. H3: Damage scales with imbalance

Stratifying by dataset imbalance ratio (Fig. 5) reveals that undersampling’s harm grows steeply with IR: on *yeast_m18* (IR= 70), ECE inflates from 0.008 to 0.395 ($\delta = +0.39$), whereas SMOTE’s inflation stays below 0.02 across the whole IR range. Discarding majority examples to balance an extremely skewed set leaves too few data to estimate probabilities, and calibration collapses.

The per-dataset view (Table III) confirms the aggregate is not an artefact of averaging: ECE rises under SMOTE on all five datasets and under undersampling on all five, while isotonic recalibration yields the lowest ECE on every dataset, including the extreme *yeast_m18*.

E. Discrimination is preserved, and the cost stays hidden

Table II makes the central irony quantitative. Minority-class F1 actually *rises* under every resampling family ($0.530 \rightarrow 0.546$ SMOTE, 0.555 RUS), and ROC-AUC moves by less

TABLE III

PER-DATASET ECE FOR KEY CONDITIONS (MEAN OVER BOTH MODELS AND TEN SEEDS). RESAMPLING RAISES ECE ON EVERY DATASET; THE EFFECT IS CATASTROPHIC FOR UNDERSAMPLING (RUS) AT HIGH IMBALANCE, WHILE ISOTONIC RECALIBRATION OF THE SMOTE MODEL RESTORES THE LOWEST ECE THROUGHOUT.

Dataset	IR	Baseline	SMOTE	RUS	SMOTE+Iso
pima	1.9	0.108	0.123	0.151	0.050
credit-g	2.3	0.093	0.100	0.181	0.043
phoneme	2.4	0.029	0.040	0.079	0.013
adult	3.2	0.022	0.024	0.123	0.011
yeast_ml8	70.1	0.008	0.019	0.395	0.006

than 0.02; PR-AUC is likewise flat (within 0.03). A practitioner tuning on any of these three metrics would conclude resampling helped, while ECE tells the opposite story—this is precisely why the cost is hidden. Notably, the calibration repair lowers F1 slightly ($\rightarrow 0.508$): post-hoc calibration optimises probability quality, not the 0.5-threshold label, so the two objectives can diverge. Reporting both is therefore essential.

F. The two ensembles start from different calibration

Averaged over conditions, the random forest is far better calibrated than histogram gradient boosting at baseline (ECE 0.023 vs 0.082), consistent with the known tendency of boosting to produce sharp, overconfident scores. Crucially, resampling degrades *both*: SMOTE raises random-forest ECE from 0.023 to 0.033 and boosting ECE from 0.082 to 0.089. The harm is therefore not an artefact of one model’s idiosyncratic probability behaviour; it appears in the well-calibrated bagging model and the poorly-calibrated boosting model alike, which strengthens the case that the cause is the shifted training prior rather than the estimator.

G. H3b: Damage grows with oversampling aggressiveness

The imbalance-ratio trend raises a follow-up question: is the harm driven by *how much* synthetic data we inject? We sweep the SMOTE oversampling ratio ρ from 0 (no oversampling) to 1.0 (full balance) on the boosting model, ten seeds per setting (Fig. 3). Aggregate ECE rises monotonically with ρ ($0.082 \rightarrow 0.084 \rightarrow 0.087 \rightarrow 0.088 \rightarrow 0.089$), while ROC-AUC is flat ($0.859 \rightarrow 0.864$) and F1 even improves slightly. The effect is small per increment but consistent in direction across datasets—phoneme, for instance, climbs steadily from 0.022 to 0.040. More synthetic minority mass means a more distorted training prior, and calibration pays for it monotonically, again with no warning from the discrimination metrics.

The extreme case makes the mechanism visible. Fig. 4 shows the reliability diagram for yeast_ml8: after undersampling rebalances a 70:1 problem to 1:1, the model’s predicted probabilities sit far above the observed frequencies—it has effectively learned the wrong prior and become severely overconfident, which is exactly the 0.395 ECE reported in Table III.

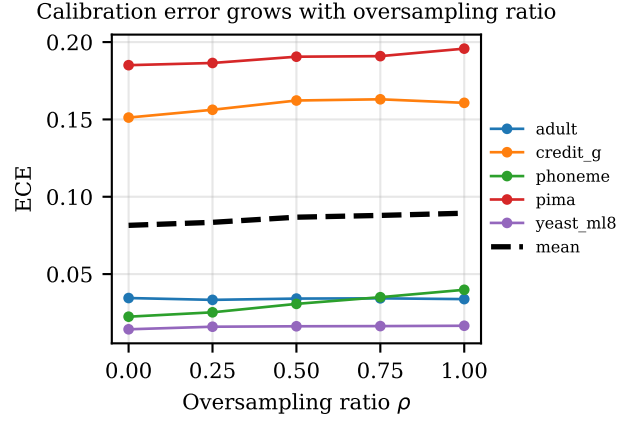


Fig. 3. ECE versus SMOTE oversampling ratio ρ (gradient boosting, ten seeds). Calibration error increases monotonically with the amount of synthetic minority data; the dashed line is the across-dataset mean.

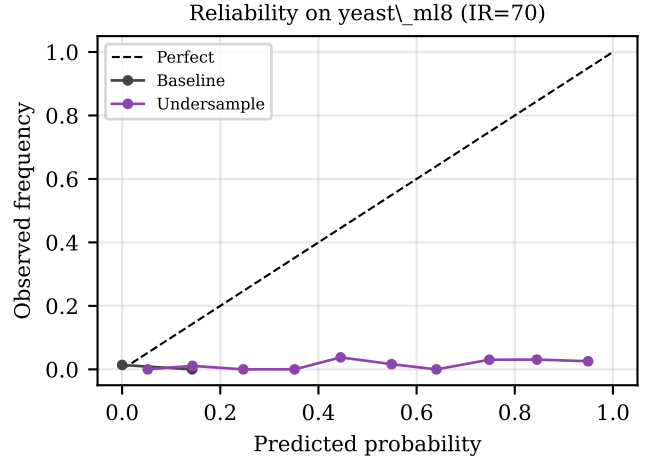


Fig. 4. Reliability diagram on yeast_ml8 (IR= 70). After undersampling (purple), predicted probabilities are grossly inflated: points fall far below the diagonal, so the model is systematically overconfident relative to the true positive rate.

H. Full per-condition breakdown

For completeness, Table IV reports ECE for all seven conditions on all five datasets. Two regularities hold without exception. First, every resampling column (SMOTE, ROS, RUS) is at least as large as the baseline on every dataset, and RUS is the largest resampler everywhere. Second, the two post-hoc-calibration columns are the smallest in every row, confirming that recalibration is a uniformly safe move for probability quality across this dataset range.

I. What resampling does not damage: feature attributions

A natural worry is that resampling might also distort *what* the model learns, not merely the scale of its probabilities. We test this with SHAP attributions on credit-g (48 features, gradient boosting), comparing mean $|\text{SHAP}|$ feature importances of the baseline and SMOTE models (Fig. 6).

TABLE IV
ECE FOR EVERY CONDITION AND DATASET (MEAN OVER BOTH MODELS, TEN SEEDS). LOWEST ECE PER ROW IN BOLD; IT IS ALWAYS A RECALIBRATED COLUMN.

Dataset	Resampling				CW bal.	Post-ho	
	Base	SMOTE	ROS	RUS		Platt	
pima	0.108	0.123	0.117	0.151	0.108	0.039	0.
credit-g	0.093	0.100	0.109	0.181	0.096	0.026	0.
phoneme	0.029	0.040	0.033	0.079	0.036	0.011	0.
adult	0.022	0.024	0.050	0.123	0.039	0.023	0.
yeast_ml8	0.008	0.019	0.010	0.395	0.009	0.004	0.

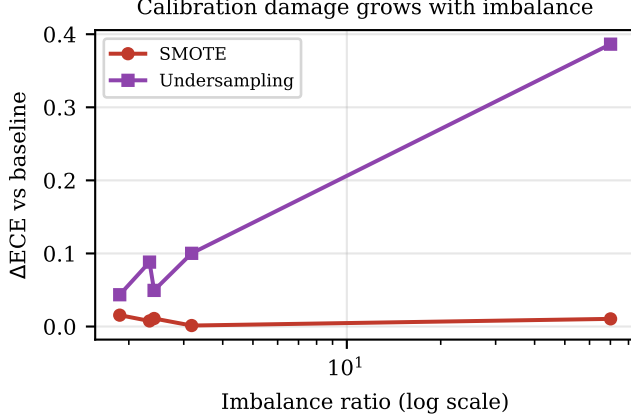


Fig. 5. ECE inflation over baseline versus imbalance ratio (log scale). Undersampling (purple) degrades sharply with imbalance; SMOTE (red) remains mild.

The two rankings are almost identical: the Spearman rank correlation between baseline and SMOTE feature importances is 0.96, and the top features (e.g. `checking_status`) keep their order and roughly their magnitudes. This is a clarifying negative result: resampling leaves the model’s learned feature structure largely intact while corrupting the probability *scale*. The damage is specific to calibration, which is exactly why a single monotone post-hoc rescaling can repair it without disturbing what the model has learned.

VI. A NEGATIVE RESULT: PRIOR CORRECTION DOES NOT TRANSFER TO SMOTE

For undersampling, the calibration damage is essentially a known prior shift, and Dal Pozzolo et al. [3] exploit this with a closed-form correction. Because SMOTE balances the training prior to $\pi_{\text{train}}=0.5$ while deployment faces the true prior π_{test} , the same analytic correction [9], [10] is tempting: it requires *no held-out calibration data* and, being monotone, cannot harm AUC,

$$p_{\text{corr}} = \frac{pr}{pr + (1-p)s}, \quad r = \frac{\pi_{\text{test}}}{\pi_{\text{train}}}, \quad s = \frac{1-\pi_{\text{test}}}{1-\pi_{\text{train}}}. \quad (2)$$

We tested this head-to-head against isotonic recalibration on the SMOTE model (Table V). The result is negative and instructive: the analytic correction does *not* improve

Feature attribution shift under SMOTE (credit_g)

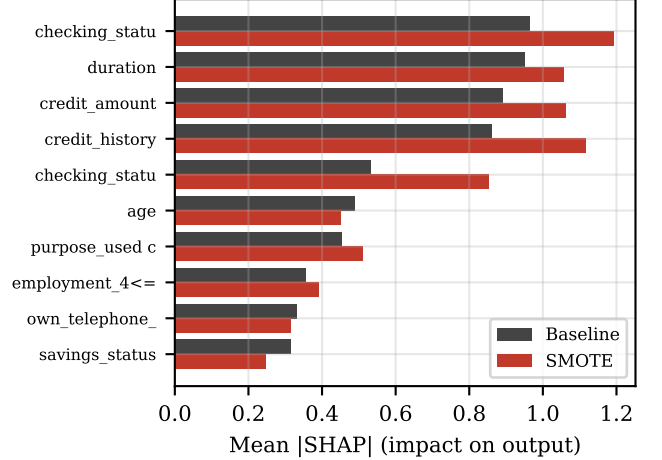


Fig. 6. Mean |SHAP| feature importance for the top-10 baseline features, baseline vs SMOTE (credit-g, gradient boosting). The attribution ranking is essentially preserved (Spearman $\rho = 0.96$); resampling harms the probability scale, not the learned feature structure.

TABLE V
ANALYTIC PRIOR CORRECTION VS. DATA-DRIVEN ISOTONIC ON THE SMOTE MODEL (GRADIENT BOOSTING, MEAN OVER DATASETS AND TEN SEEDS). PRIOR CORRECTION PRESERVES AUC BUT FAILS TO IMPROVE ECE; ONLY ISOTONIC REPAIRS CALIBRATION.

Method	ECE	AUC	F1
SMOTE (uncalibrated)	0.090	0.863	0.550
+ analytic prior correction	0.096	0.863	0.512
+ isotonic (data-driven)	0.026	0.835	0.507

calibration on SMOTE (ECE 0.090 \rightarrow 0.096, Wilcoxon $p=0.80$ —no significant change), and on `adult` it is markedly worse (0.035 \rightarrow 0.082). Data-driven isotonic recalibration, by contrast, cuts ECE sharply (0.090 \rightarrow 0.026, $p<10^{-14}$). As predicted, prior correction leaves AUC untouched (0.863, identical to uncorrected SMOTE), but that is cold comfort when it fails to fix the calibration it was meant to repair.

The explanation is that the correction’s assumption is violated. Prior adjustment is exact only when resampling shifts $p(y)$ alone, leaving the class-conditional density $p(\mathbf{x} | y)$ intact—true for random undersampling, but *not* for SMOTE, which synthesises new minority points by interpolation and thereby alters $p(\mathbf{x} | y=1)$ itself. The shift SMOTE induces is therefore not a pure prior shift, and no prior-only formula can undo it. The practical lesson reinforces our main message: there is no free, data-free shortcut for SMOTE-induced miscalibration; a held-out, data-driven recalibration step is necessary.

VII. DISCUSSION

Our results reframe a routine modelling choice. Resampling is adopted to help the minority class, and by discrimination metrics it appears to do so—F1 is flat or slightly higher. But the same intervention quietly corrupts the predicted probabilities, and the corruption is largest exactly where imbalance is

most severe and resampling is most tempting. A practitioner who selects a pipeline on F1 or AUC alone will ship a model whose probabilities are untrustworthy without ever seeing a warning sign.

The remedy is cheap and well understood: recalibrate after resampling. One Platt or isotonic step restored calibration below baseline in every aggregate we measured, and because both are monotone they leave the decision ranking intact. The practical guidance is therefore simple—if you resample and your downstream system consumes probabilities, add a held-out recalibration stage, and prefer SMOTE or oversampling over aggressive undersampling when imbalance is extreme.

A. Practical recommendations

We distil the findings into four concrete guidelines. **(1) Always report calibration.** ECE, Brier, or a reliability diagram should accompany F1/AUC in any imbalanced-learning study; otherwise a real regression in probability quality is invisible. **(2) Recalibrate after resampling** on a held-out split whenever probabilities feed a threshold or expected-cost decision; one isotonic or Platt step suffices and restored ECE below baseline in every case we measured. **(3) Avoid aggressive undersampling at high imbalance.** Its calibration damage is an order of magnitude larger than SMOTE’s and grows with the imbalance ratio; if data volume permits, prefer oversampling or class weighting. **(4) Prefer class weighting when a calibrated probability is the goal and the model supports it**, since it barely perturbed calibration (ECE $0.052 \rightarrow 0.058$) while still addressing the loss imbalance.

B. Reproducibility

All datasets are public OpenML sets identified by numeric ID; the preprocessing (one-hot encoding, median imputation, stratified subsampling of `adult` to 8,000 rows) is deterministic given the seed. Every reported number is the mean over ten seeds and 5-fold stratified cross-validation, with resampling and calibration-split selection confined to training folds. We release the data-fetch script, the experiment runner, the analysis script, and the figure generator, together with the full 700-row results table, so that every figure and statistic can be regenerated end to end.

VIII. LIMITATIONS

Our scope is deliberately bounded and our claims should not be overstated. (i) We study two tree ensembles on tabular data; neural models and other modalities may behave differently. (ii) The calibration–AUC trade-off we report is confounded with the size of the held-out calibration split; a study that varies that split would isolate the two effects. (iii) ECE with fixed-width bins is a known imperfect estimator; we mitigate by also reporting Brier score and reliability diagrams, which agree. (iv) Five datasets, though spanning a wide imbalance range, cannot represent every regime. These bounds point to clear follow-up work rather than undermining the central, robustly significant finding.

IX. CONCLUSION

Resampling for class imbalance carries a hidden cost: it degrades the probability calibration of tree ensembles, significantly and—for undersampling under heavy imbalance—severely, while leaving the discrimination metrics that practitioners watch unchanged. A single post-hoc recalibration step repairs the damage at a negligible cost to ranking. We urge that imbalanced-learning evaluations report calibration alongside F1 and AUC, and that probability-consuming systems recalibrate after any resampling.

REFERENCES

- [1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [2] A. Fernández, S. García, F. Herrera, and N. V. Chawla, “SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.
- [3] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in *Proc. IEEE Symp. Series on Computational Intelligence (SSCI)*, 2015, pp. 159–166.
- [4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [6] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, 2005, pp. 625–632.
- [7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. 34th Int. Conf. Machine Learning (ICML)*, ser. PMLR, vol. 70, 2017, pp. 1321–1330, arXiv:1706.04599.
- [8] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, “An experimental investigation of calibration techniques for imbalanced data,” *IEEE Access*, vol. 8, pp. 127 343–127 352, 2020.
- [9] M. Saerens, P. Latinne, and C. Decaestecker, “Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure,” *Neural Computation*, vol. 14, no. 1, pp. 21–41, 2002.
- [10] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2001, pp. 973–978.