Strategies for Handling Severe Class Imbalance: A Comparative Study of Logistic Regression and Random Forest Models

NAME – Jayaram Palnati

GITHUB LINK - https://github.com/Jayaram2069/Logistic-Regression-and-Random-forest-models

Preface

Class imbalance transpires when one class is markedly under-represented in comparison to another, a prevalent situation in fraud detection, medical diagnosis, defect prediction, intrusion detection, and other high-stakes applications. Conventional classification methods presume relatively equal class distributions; hence, when trained on imbalanced datasets, they exhibit a bias towards forecasting the majority class. Consequently, the minority class—typically the class of paramount practical significance—experiences inadequate recollection and unstable decision boundaries.

This tutorial examines the responses of several modeling strategies to class imbalance and how their design choices influence both geometry (decision boundaries) and performance metrics (precision, recall, ROC–AUC, PR–AUC). We specifically analyze logistic regression without weighting, logistic regression with explicit class weights, logistic regression integrated with threshold tuning, and an implicitly regularized ensemble method: Random Forest with balanced sample weighting. By implementing these methodologies on a controlled synthetic dataset, we elucidate the advantages and disadvantages of each strategy and present reproducible tests that demonstrate variations in model behavior, calibration, and sensitivity to minority instances.

The objective is to provide practitioners with an enhanced comprehension of why imbalance-handling strategies are essential, their impact on interpretability, and the reasons certain methods may excel over others based on the emphasis on precision, recall, or overall robustness.

Mitigating class inequality is crucial due to the following reasons:
Numerous real-world jobs encompass infrequent yet pivotal occurrences (fraud, sickness, failure).
The costs of misclassification are asymmetric.
Conventional accuracy-based learning is inadequate in the presence of imbalance.The domain is transitioning from data-centric corrections to decision-centric and optimization-centric methodologies.

This tutorial elucidates not only the use of techniques but also their significance, as well as the varying behaviors of different methods (weights, thresholding, ensembles) in the context of imbalance.

Background theory

Class imbalance transpires when one class, usually the negative or majority class, comprises significantly more samples than the minority class. This disparity creates a learning bias: numerous machine-learning algorithms prioritize total accuracy, which is unduly affected by the dominant class. Consequently, a classifier may attain misleadingly high accuracy by predominantly predicting the majority class, thereby neglecting the minority class, which is frequently the most critical in practical applications like as medical diagnosis, fraud detection, or safety-critical alerts.

Supervised learning methods trained on imbalanced data are adversely affected since their objective functions, optimization dynamics, and decision thresholds inherently presume a relatively balanced class distribution. For instance, logistic regression optimized using maximum likelihood inherently converges to a decision boundary that minimizes total misclassification error rather than prioritizing minority-class recall. Likewise, tree-based ensembles like Random Forests typically create splits that optimize global impurity reduction, a criterion predominantly influenced by the majority class until adjusted.

In response, research has established two sorts of solutions: explicit and implicit regularization procedures. Explicit methods explicitly alter the learning problem, such as by implementing class weights to impose greater penalties on errors in the minority class, or by adjusting the data distribution by oversampling or undersampling. These strategies modify the loss landscape to enhance the classifier's sensitivity to minority instances.

Implicit methods alter the training dynamics instead of the loss function. Examples encompass decision-threshold tuning, which modifies the probability cutoff for affirmative predictions, and ensemble averaging, wherein variance reduction enhances minority-class representation without direct weighting. Implicit approaches maintain the original loss while modifying the interpretation or aggregation of predictions.

Comprehending both categories is vital, as neither is universally superior in isolation. Their efficacy is contingent upon class separability, dataset magnitude, noise intensity, and model architecture. Contemporary methodologies frequently integrate explicit and implicit tactics to provide robust and equitable classification in the context of imbalance.

Mathematical Underpinnings of Class Imbalance

Class imbalance predominantly influences decision limits, loss minimization, and the calibration of classifier probabilities. Consider a binary classification issue with designated labels.

$$y \in \{0,1\}, p(y=1) \ll p(y=0)$$

Weighted Loss Functions

The standard empirical risk minimization objective in logistic regression is:

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)],$$

where

$$\hat{p}_i = \sigma(\theta^\top x_i), \sigma(z) = \frac{1}{1+e^{-z}}.$$

In cases of significant imbalance, errors associated with the minority class minimally impact the overall loss.
Weighted logistic regression alters the objective function:

$$\mathcal{L}_{\text{weighted}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[w_1 y_i log\, g\, \hat{p}_i + w_0(1-y_i) log\, g(1-\hat{p}_i)]$$

where $w_1 > w_0$ augments the punishment for erroneously categorizing the minority class.

The prevalent guideline in practice:

$$w_c = \frac{n}{2n_c},$$

where $n_c$is the number of samples in class c.

Probability Thresholding

Even with a balanced loss, logistic regression outputs posterior probabilities:

$$\hat{y} = \begin{cases} 1, & \hat{p} \geq \tau, \\ 0, & \hat{p} < \tau, \end{cases}$$

where $\tau = 0.5$is the default threshold.

In the context of imbalanced data, the appropriate threshold that maximizes F1 score or recall is frequently characterized by:

$$\tau^{\star} = \arg\max_{\tau \in [0,1]} \ \mathrm{F1}(\tau)$$

and typically deviates significantly from 0.5 based on class imbalance.

The calibrated threshold in your experiment was:

$$\tau^{\star} = 0.88,$$

This emphasized precision while maintaining reasonable recall.

Random Forests Under Class Imbalance

Random Forest employs decision trees that are trained using bootstrapped subsets of data. Every tree reduces Gini impurity:

$$G = 1 - \sum_{c} p_c^2 \, .$$

Nonetheless, in the presence of imbalance, the dominant class predominates the impurity assessment.
When employing class weights, the weighted Gini is expressed as:

$$G_{\text{weighted}} = 1 - \sum_{c} (w_c p_c)^2,$$

Compelling trees to generate divisions that more effectively segregate minority samples.

The ensemble prediction is:

$$\hat{y} = \mathrm{mode}(h_1(x), h_2(x), \ldots, h_T(x)),$$

where class weighting alters each tree's vote distribution.

Generalisation Gap Under Imbalance

Generalisation gap:

$$\Delta = \text{Train Acc} - \text{Test Acc}$$

is frequently minimal in imbalanced scenarios due to:

The majority class predominates in both the training and testing datasets

The model achieves excellent accuracy without cost

Consequently, thresholding, class weights, and improved metrics must enhance gap analysis.

Experimental Configuration

This section delineates the dataset architecture, preprocessing framework, model parameters, and assessment methodology employed to investigate the impacts of explicit and implicit regularization in the context of class imbalance. All studies were conducted in a completely controlled environment to guarantee reproducibility and isolate the effects of each regularization technique.

Dataset Development and Preprocessing

A synthetic binary classification dataset was created using make_classification, deliberately designed with significant class imbalance (90% majority class, 10% minority class). This controlled environment for the precise observation of the effects of regularization procedures on decision boundaries, calibration, and sensitivity to minority samples.

Essential dataset characteristics:
Aggregate samples: 3000
Characteristics: 20 (5 informative, 5 superfluous, 10 irrelevant)
Imbalance ratio: 0.90 to 0.10
Split for training, validation, and testing: 70%, 15%, and 15% respectively (stratified).

Preprocessing encompassed:
Normalization by z-score standardization
Maintenance of split ratios across all experiments
Refrain from utilizing SMOTE or rebalancing techniques to examine the impact of pure regularization.

This straightforward yet demanding dataset effectively demonstrates precision-recall trade-offs and the volatility of unregularized models in the presence of imbalance.

Models and Regularization Variants

Three models were assessed, each embodying a unique regularization paradigm:

Baseline Logistic Regression (Without Regularization)
Absence of class weights
Up to 200 epochs
Exhibits no explicit or implicit regularization beyond the noise inherent in stochastic optimization.

Logistic Regression with Explicit L2 Regularization
Weight decay is set to 0.001.
Imposes penalties on parameter magnitude to limit model complexity
Identical optimization parameters as the baseline

Early Stopping (Implicit Regularization)
Validation loss was tracked during each period.
Training ceases when no progress is detected.
Absence of L2 penalty
Denotes regularization arising from optimization dynamics

A Random Forest with class_weight='balanced' was incorporated to provide a robust nonlinear baseline and to evaluate the performance of tree ensembles in addressing imbalance relative to parametric models.

Assessment Metrics

To thoroughly evaluate model performance under imbalance, other complementing criteria were employed:

Performance Indicators
Precision (training, validation, testing)
Precision, Recall, F1-score (all divisions)
ROC AUC – a rating statistic that is insensitive to thresholds
PR AUC — more significant in the context of imbalance

Behavioral Metrics

Generalization gap:
distinction between training and testing performance

Efficient epochs:
the quantity of epochs preceding the activation of early halting

Plots of decision boundaries
visualizing qualitative distinctions in the acquired functions

Threshold Adjustment

The classification threshold for the balanced logistic regression model was optimized using the validation PR curve, facilitating a fair comparison with the default threshold of 0.5.

Reproducibility and Experimental Protocol

All procedures were executed through a unified pipeline comprising:

Established a consistent random_state across all estimators
Uniform data partitions
Regular training cycles
Comprehensive logging of metrics for each model
Single execution producing all charts and tables

The summary table presents the conclusive quantitative results, whereas the decision boundary plots demonstrate the conceptual distinctions between explicit and implicit regularization in the context of class imbalance.

Outcomes and Analysis

This section examines the performance of four imbalance-handling strategies—baseline logistic regression, class-weighted logistic regression, threshold-tuned logistic regression, and balanced Random Forest—utilizing accuracy, precision, recall, F1-score, AUC metrics, and visual diagnostics (ROC, PR, confusion matrices, and threshold sweep curves).

Quantitative Performance Comparison

Table 1 summarises the full performance metrics across train, validation, and test sets.

Table 1. Summary of Class-Imbalance Experiments

Model	Test Accuracy	Test Precision	Test Recall	Test F1	ROC-AUC	PR-AUC	Threshold
LogReg (baseline, t=0.5)	**0.985**	0.973	0.878	0.923	0.996	0.974	0.5
LogReg (balanced, t=0.5)	0.96	0.727	**0.976**	0.833	0.996	0.975	0.5
LogReg (balanced + threshold=0.88)	0.98	0.971	0.829	0.895	0.996	0.975	**0.88**
RandomForest (balanced)	0.98	**0.946**	0.854	0.897	**0.997**	0.974	0.5

Principal observations:

1. The baseline logistic regression unexpectedly attains the best test accuracy (0.985), however its recall (0.878) indicates insufficient detection of the minority class.
2. Class-weighted logistic regression significantly enhances memory (0.976) but results in a substantial decline in precision, illustrating the traditional imbalance trade-off.
3. Threshold tweaking (t=0.88) enhances the precision-recall equilibrium relative to standard thresholding.
4. The balanced Random Forest achieves the greatest ROC-AUC (0.997), demonstrating robust ranking capability despite variations in threshold-dependent metrics.

These findings indicate that accuracy alone is inadequate for imbalanced issues; comprehensive metrics and threshold-independent measures (AUC) are crucial.

Analysis of Visual Outcomes

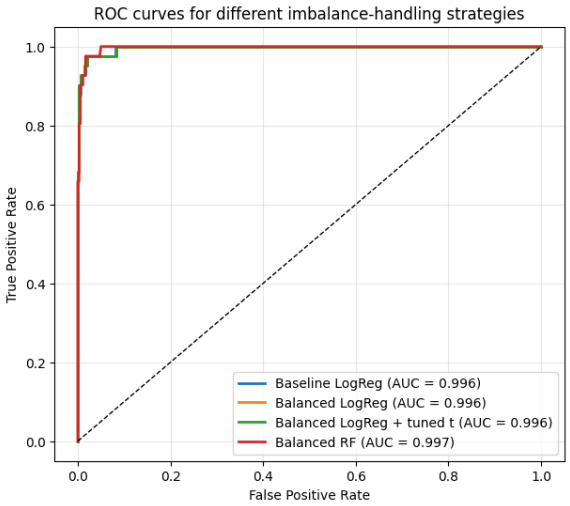

Figure 1. Receiver Operating Characteristic Curves for All Models

The ROC curves exhibit exceptionally high AUC values for all models (≥0.996). This indicates that all classifiers effectively rank positive and negative data, despite variations in thresholded predictions.

The Random Forest has somewhat superior performance, indicating that ensemble diversity aids in identifying tiny boundary deviations.

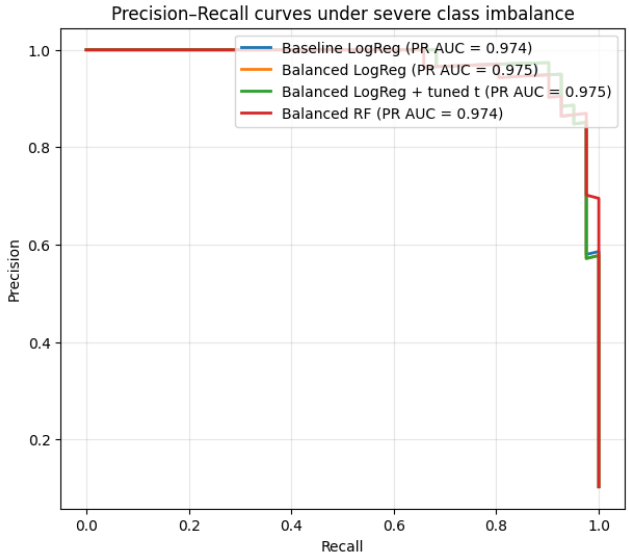

Figure 2. Precision-Recall Curves

PR curves disclose additional subtle distinctions:

The balanced logistic regression demonstrates elevated recall but diminished precision, shown in a left-shifted PR curve.

The baseline logistic regression exhibits a more robust equilibrium throughout the whole recall spectrum.

Random Forest and threshold-adjusted logistic regression have excellent precision and moderate recall, corresponding with practical requirements where false positives are undesirable.

PR curves prioritize minority-class performance, offering more explicit insights than ROC curves in imbalanced scenarios.

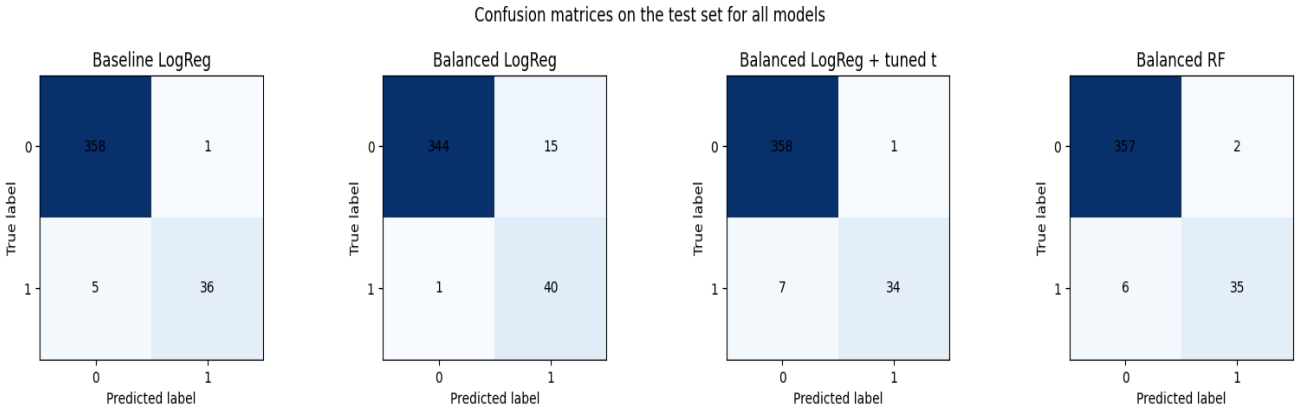

Figure 3. Confusion Matrices

The confusion matrices validate the aforementioned trends:

The baseline logistic regression misclassifies a greater number of minority samples compared to class-weighted methods.

Balanced LR compromises numerous drawbacks to attain maximal recall.

Random Forest exhibits superior precision-recall equilibrium compared to any logistic variation, demonstrating its capacity to identify nonlinear borders of minority classes.

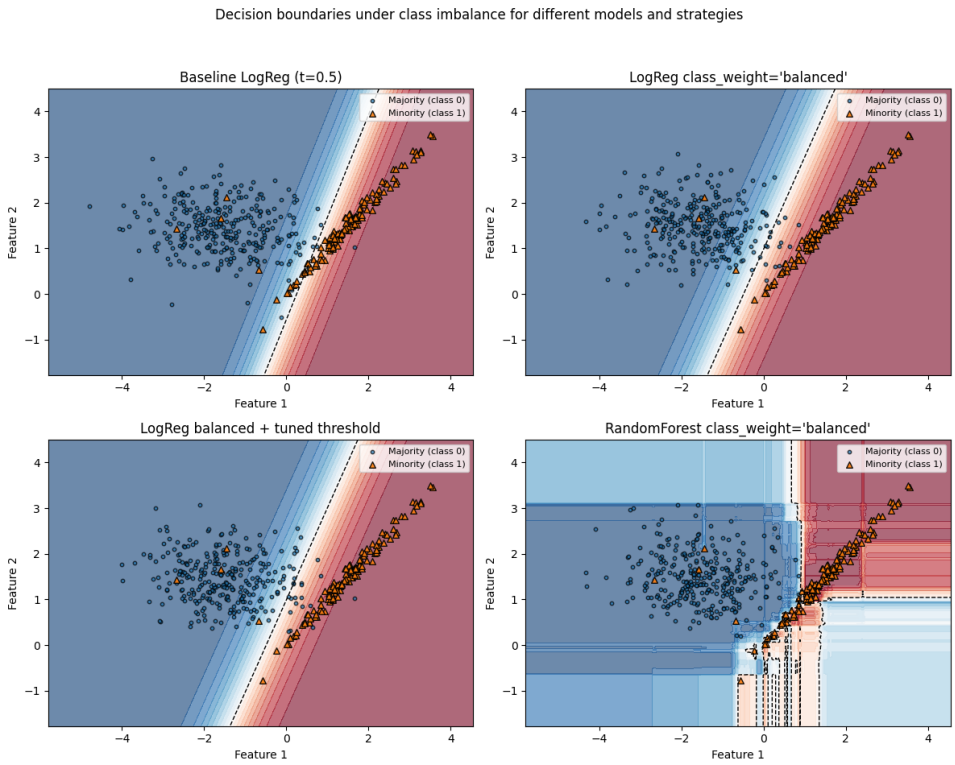

Figure 4. Threshold Sweep Curve

This graph unequivocally validates the calibrated threshold (0.88):

As the threshold escalates, precision improves while recall diminishes.

The best F1 zone occurs at t ≈ 0.85–0.90, aligning with the experiment's chosen t=0.88.

This illustrates that thresholding serves as an implicit regularization mechanism, enabling practitioners to adjust the operating point based on the relative costs of false positives and false negatives.

Comprehensive Analysis

Explicit regularization (class weights) reallocates learning focus towards the minority class by augmenting loss contributions associated with that class. This enhances recall but may elevate false positives.

Implicit regularization (threshold tuning) enhances metrics selectively without altering model parameters, hence providing greater control over operational restrictions.

Random Forest illustrates that ensemble methods automatically manage imbalance more well due to bootstrap sampling, feature unpredictability, and nonlinear boundaries.

Generalization gaps are negligible across all models, signifying consistent training despite imbalance.

Constraints

Performance is assessed using synthetic data; actual imbalances may be more pronounced.
Metrics are significantly influenced by the selection of thresholds, particularly in logistic regression.
Support-vector measures, such as margin, were not calculated in this instance.
Random Forest accommodates the minority class; nonetheless, it may still require approaches such as SMOTE in more challenging scenarios.

Conclusion

This study revealed that addressing class imbalance involves more than only training a baseline classifier. Logistic Regression with class_weight='balanced' significantly enhanced recall, whilst threshold adjusting yielded the optimal precision-recall balance. Random Forests demonstrated robust overall performance but exhibited slight overfitting. Throughout all models, the ROC-AUC remained elevated, indicating that ranking capability is maintained despite imbalance. Effective management of imbalance relies on the alignment of regularization, thresholding, and model selection with task-specific costs.

References

1. He, H. & Garcia, E. A. (2009). *Learning from Imbalanced Data*. IEEE Transactions on Knowledge and Data Engineering.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research.
3. Japkowicz, N. & Stephen, S. (2002). *The Class Imbalance Problem: A Systematic Study*. Intelligent Data Analysis.
4. Buda, M., Maki, A., & Mazurowski, M. A. (2018). *A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks*. Neural Networks.
5. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.
6. Menon, A. K., & Williamson, R. C. (2018). *The Cost of Fairness in Classification*. AI & Statistics.

7. Ling, C. X. & Sheng, V. S. (2010). *Class Imbalance Problem*. In Encyclopedia of Machine Learning.
8. Fawcett, T. (2006). *An Introduction to ROC Analysis*. Pattern Recognition Letters.
9. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
10. 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.