Practical Project
Machine Learning I – University of Porto
Academic Year 2024/2025

Binary Classification

Under Class Imbalance

Guilherme Batista (201910641)
Yan Coelho (202300916)

EXECUTIVE SUMMARY

- Goal: Build and evaluate a binary classifier implemented 100% in pure Python (no scikit-learn) and make it imbalance-aware.

- Approach:
  - Robust Data Handling: Automated cleaning, encoding, missing value imputation, and target binarization
  - Stratified Train/Test Split: Preserves class distribution for fair evaluation.
  - Feature Scaling: StandardScaler normalization applied across datasets
  - 5 Custom Models Evaluated (Logistic Regression and its variances)
  - Performance Metrics: Accuracy, Balanced Accuracy, Precision, Recall, F1, ROC-AUC, G-Mean
  - Visual Diagnostics: ROC and Confusion Matrices for all models
  - Aggregate Evaluation: Metric summary tables (mean ± std) across datasets, best values highlighted
  - Statistical Analysis: Paired t-tests with confidence intervals vs. baseline (BCE), identifying significant improvements

EXECUTIVE SUMMARY

- Summary of Results:: Build and evaluate a binary classifier implemented 100% in pure Python (no scikit-learn) and make it imbalance-aware.

Model	accuracy	balanced_accuracy	precision	recall	f1	auc	gmean
BCE	0.929 ± 0.056	0.709 ± 0.183	0.734 ± 0.307	0.688 ± 0.375	0.688 ± 0.341	0.874 ± 0.139	0.531 ± 0.376
BCE New Sigmoid	0.925 ± 0.070	0.742 ± 0.199	0.702 ± 0.351	0.725 ± 0.373	0.700 ± 0.361	0.861 ± 0.163	0.579 ± 0.399
Focal	0.918 ± 0.067	0.724 ± 0.191	0.691 ± 0.300	0.749 ± 0.350	0.701 ± 0.317	0.874 ± 0.140	0.561 ± 0.380
Focal Dynamic Alpha	0.641 ± 0.284	0.712 ± 0.149	0.453 ± 0.355	**0.969 ± 0.074**	0.538 ± 0.324	0.876 ± 0.137	0.584 ± 0.313
Weighted	0.701 ± 0.221	0.764 ± 0.142	0.490 ± 0.350	0.921 ± 0.125	0.560 ± 0.298	0.872 ± 0.140	0.704 ± 0.240
Weighted_New_Sigmoid	0.837 ± 0.178	**0.796 ± 0.175**	0.651 ± 0.341	0.837 ± 0.214	0.690 ± 0.302	0.865 ± 0.152	**0.760 ± 0.234**

- Highlighted illustrate the entries that achieves the highest (and statistically-significant) value (paired Wilcoxon, *p* < 0.05).

SELECTED ALGORITHM AND DATA CHARACTERISTIC

Logistic Regression

- Why? Logistic regression is interpretable, probabilistic, and efficient, making it a strong baseline for binary classification, especially in imbalanced settings where probability calibration and adaptation via weighted losses can mitigate imbalance.

- Compared to Random Forest and SVM, it offers faster training, better calibration, and easier interpretation.

- While SVMs are powerful with kernels and perform well on high-dimensional data, they lack scalability and require extra steps for probability estimates — making logistic regression more practical and reliable in many real-world, imbalance-sensitive applications.

SELECTED ALGORITHM AND DATA CHARACTERISTIC

Behavior of the algorithm in a highly imbalanced dataset

- The binary cross-entropy (BCE) loss optimizes overall accuracy, not class balance.
- Tends to be biased toward the majority class if it's overrepresented (high accuracy, low recall for minority).
- High accuracy, poor F1, recall.
- Underestimates the importance of rare events — critical in fraud detection, medical diagnosis, etc.

Proposal

- Motivation: Recover the model's ability to correctly detect minority class instances without significantly compromising the overall performance.
- Proposed Modifications to Logistic Regression:
  - Weighted BCE Loss: model will be penalized more for misclassifying minority class examples
    - Adds class weights to penalize errors on the minority class more heavily
    - Helps the model detect rare cases more reliably
  - Custom Sigmoid Function:
    - Improves learning for borderline samples, often belonging to the minority class
  - Focal Loss:
    - Reduces the loss contribution of majority examples
    - Focuses the model on hard-to-classify instances
  - Dynamic Focal Loss:
    - Adaptively adjusts how much to focus on hard examples during training
    - Balances learning between stability and focus over time

Proposal

Logic of each model modification

- Standard Binary Cross-Entropy (BCE): $-\log(p)$
  p: predicted probability for the correct class.

- Focal Loss: $-\alpha(1-p)^{\gamma} * \log(p)$
  α: balancing fator to rebalance class weights
  $\gamma \geq 0$ is the focusing parameter (if equals 0 is standard BCE)

- Weighted BCE (Class-Weighted Loss): $-\alpha \cdot y \cdot \log(p) - (1-\alpha)(1-y) \cdot \log(1-p)$
  α: weighting factor; p: predicted probability; y: true label

- Custom Sigmoid Transformation: $1/1+e^{(-\alpha z)}$
  α: steepness control (standard when equals 1)

Empirical Study

experimental setup

1) Robust preprocessing
    1) placeholder and string-“NaN” cleaning;
    2) per-column type coercion;
    3) label-encoding of categorical/boolean values;
    4) mean imputation for numerical NaNs;
    5) complete-case deletion of columns (all-NaN) and rows (remaining NaNs);
2) Target binarization;
3) Data partitioning;
4) Feature scaling;
5) Model training;
6) Prediction & metric collection;
7) Visual diagnostics;
8) Synthesis of metrics and paired Student t-tests (95% CI) between each model and BCE base

Empirical Study

datasets and their characteristics

- 50 imbalanced datasets with different scenarios
- heterogeneous mixes of numeric, categorical and boolean variables and with missing data
- datasets differ in imbalance ratio, feature dimensionality and sample size, enabling the paired tests to gauge the robustness of each loss variant across a spectrum of imbalance severities
- Stratified Split to guarantee the same class proportion in train/test;

Empirical Study

hyperparameters of the algorithm

- Dynamic Grid search executed as an attempt to find the best combination

- End result:
  - Learning Rate: 0.05, Convergence speed vs. stability
  - Penalty: L2, Generalisation and numeric robustness
  - Tolerance:1e-6, Solution accuracy
  - max_iters: 1000 Runtime control & fairnes

Empirical Study

performance estimation methodology within dataset

- Stratification – guarantees that both train and test contain the same minority-class prevalence.
- Single split. High variance due different datasets is later averaged out across the portfolio of datasets.
- Scaling in the pipeline, so the model would behave on brand-new, unseen data.
- Metrics:
  - Accuracy – irrelevant to the imbalance problem.
  - Balanced accuracy – mean of class-wise recalls, insensitive to skew.
  - Recall – sensitivity of the minority class, the chief target metric.
  - Precision – proportion of predicted positives that are correct.
  - F1 – harmonic mean of precision/recall.
  - g-mean – geometric mean of class-wise recalls, recommended for imbalance
  - AUC – threshold-independent discrimination.
- Diagnostic plots (ROC, confusion matrix) are produced.

Empirical Study

performance estimation methodology across dataset synthesis

- Metrics: Accuracy, Balanced Accuracy, Precision, Recall, F-score, ROC-AUC, G-mean

- For every model and metric, the script stores mean ± standard deviation of the metrics score from each dataset and average it to report final pipeline perfomance.

- Paired two-tailed t-tests comparing each model variant to the BCE baseline (95 % CI) in metrics focused on imbalanced-oriented metrics: F1, balanced accuracy, recall, G-mean

- Fairness – All models share the same split, preprocessing, hyper-parameters and stopping conditions, so observed differences trace back to loss design, not to incidental implementation choices.

Analysis of Results

metrics

=== Best values with statistical significance highlighted ===

Model	accuracy	balanced_accuracy	precision	recall	f1	auc	gmean
BCE	0.929 ± 0.056	0.709 ± 0.183	0.734 ± 0.307	0.688 ± 0.375	0.688 ± 0.341	0.874 ± 0.139	0.531 ± 0.376
BCE New Sigmoid	0.925 ± 0.070	0.742 ± 0.199	0.702 ± 0.351	0.725 ± 0.373	0.700 ± 0.361	0.861 ± 0.163	0.579 ± 0.399
Focal	0.918 ± 0.067	0.724 ± 0.191	0.691 ± 0.300	0.749 ± 0.350	0.701 ± 0.317	0.874 ± 0.140	0.561 ± 0.380
Focal Dynamic Alpha	0.641 ± 0.284	0.712 ± 0.149	0.453 ± 0.355	**0.969 ± 0.074**	0.538 ± 0.324	0.876 ± 0.137	0.584 ± 0.313
Weighted	0.701 ± 0.221	0.764 ± 0.142	0.490 ± 0.350	0.921 ± 0.125	0.560 ± 0.298	0.872 ± 0.140	0.704 ± 0.240
Weighted_New_Sigmoid	0.837 ± 0.178	**0.796 ± 0.175**	0.651 ± 0.341	0.837 ± 0.214	0.690 ± 0.302	0.865 ± 0.152	**0.760 ± 0.234**

BCE: High accuracy and precision but misses 31% of rare cases, g-mean confirm minority low perfomance

BCE NS: overall improvement from base but precision lowers.

Focal: higher recall for minority cases, also more false positives (lower prec.)

Focal DA: high recall, i.e., great if missing one is unacceptable due, bad if false alarm is costly

Weighted: a milder version of dynamic focal: good safety-net, noisy alarms

Weighted NS: overall balance is most favourable.

Analysis of Results

metrics with t-test

=== Best values with statistical significance highlighted ===

Model	accuracy	balanced_accuracy	precision	recall	f1	auc	gmean
BCE	0.929 ± 0.056	0.709 ± 0.183	0.734 ± 0.307	0.688 ± 0.375	0.688 ± 0.341	0.874 ± 0.139	0.531 ± 0.376
BCE New Sigmoid	0.925 ± 0.070	0.742 ± 0.199	0.702 ± 0.351	0.725 ± 0.373	0.700 ± 0.361	0.861 ± 0.163	0.579 ± 0.399
Focal	0.918 ± 0.067	0.724 ± 0.191	0.691 ± 0.300	0.749 ± 0.350	0.701 ± 0.317	0.874 ± 0.140	0.561 ± 0.380
Focal Dynamic Alpha	0.641 ± 0.284	0.712 ± 0.149	0.453 ± 0.355	**0.969 ± 0.074**	0.538 ± 0.324	0.876 ± 0.137	0.584 ± 0.313
Weighted	0.701 ± 0.221	0.764 ± 0.142	0.490 ± 0.350	0.921 ± 0.125	0.560 ± 0.298	0.872 ± 0.140	0.704 ± 0.240
Weighted_New_Sigmoid	0.837 ± 0.178	**0.796 ± 0.175**	0.651 ± 0.341	0.837 ± 0.214	0.690 ± 0.302	0.865 ± 0.152	**0.760 ± 0.234**

Highlighted: Statistical significance ($p<0.05$) against base BCE

Need highest few misses of the rare (recall)?

Focal Dynamic Alpha (0.969).

Need highest few false alarms (precision/accuracy)?

BCE baseline (0.734 / 0.929).

Want most balanced performance?

Weighted New Sigmoid

Care about False alarms and few misses of rare (F1)?

Focal

Conclusions

- Standard BCE achieved the highest accuracy and precision but showed bias toward the majority class.
- Weighted BCE and Focal Loss indicated better sensivity to minority class
- Weighted_New_Sigmoid achieved the best balanced accuracy and G-Mean, combining benefits of reweighting and decision-boundary sharpening.
- Focal Dynamic Alpha improved recall greatly but suffered from instability in F1 and accuracy, indicating trade-offs.
- Statistical significance confirmed several improvements over BCE baseline, especially in recall and balanced metrics.

Future Work

- Apply PR-AUC to better recognize performance regarding imbalanced class.
- Experiment replacing the single split by k-fold cross validation;

THANK YOU!

FOR YOUR ATTENTION