Practical Assignment ML1

Project made by:
Mariana Fonseca 202303686
Beatriz Seabra 202303729

...

Logistic Regression

on Imbalanced Data

Executive Summary

Step 1
Project Overview

Step 2
Import Required Libraries

Step 3
Baseline Model: Logistic Regression Implementation

Step 4
Data Loading and Preprocessing

Step 5
Baseline Model Training and Evaluation

Step 6
Focal Loss Variant: Implementation and Evaluation

Step 7
Comparative Analysis: Metrics and Visualizations

Step 8
Advanced Visualizations and Statistical Analysis

Step 9
Conclusions and References

1. Project Overview

We implement a binary Logistic Regression classifier from scratch and evaluate its performance on an imbalanced benchmark dataset. We then propose and implement a variant to mitigate class imbalance.

Goals

01 **Thoroughly evaluate the baseline Logistic Regression model**

Evaluate its performance and limitations on real, imbalanced datasets using multiple metrics.

02 **Address imbalanced classification**

Highlight why standard models fail for minority class detection and why accuracy is misleading.

Selected algorithm and data characteristic

Description

Selected Algorithm: Logistic Regression

Logistic Regression is a linear classification algorithm that models the probability of a binary outcome using the sigmoid function applied to a linear combination of input features. It is trained by minimizing the binary cross-entropy loss via gradient descent. In this project, we implemented Logistic Regression from scratch, allowing full control over the optimization process and loss function.

Data Characteristic: Imbalanced Datasets

The datasets used in this project are highly imbalanced, meaning the minority class (the class of interest) is significantly underrepresented compared to the majority class. In most datasets, the minority class accounts for less than 20% of the samples, and in some cases, it is almost negligible.

Selected algorithm and data characteristic

Discussion of the behavior of the algorithm concerning the selected data characteristic

Baseline Logistic Regression

Tends to be biased toward the majority class, as it optimizes overall accuracy. Often fails to detect minority class samples, resulting in high accuracy but very low recall and F1-score for the minority class. The model's decision boundary is influenced by the dominant class, leading to poor generalization for rare events.

Observed Effects in Results

High accuracy is misleading, as the model simply predicts the majority class most of the time. Minority class metrics (recall, F1) are often near zero, confirming the algorithm's inability to handle imbalance. Performance is highly variable across datasets, depending on the degree of imbalance.

```python
# Standard imports
import os
import zipfile
import glob
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import f1_score, precision_recall_curve, auc, classification_report
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import accuracy_score, recall_score, roc_auc_score, f1_score, classification_repor

from collections import Counter

# Plot styling
plt.rcParams.update({'figure.figsize': (8, 6), 'axes.grid': True})

# Set random seed for reproducibility
RANDOM_STATE = 42
np.random.seed(RANDOM_STATE)

# Numerical stability constant for losses
EPS = 1e-15
```

2. Import Required Libraries

All necessary Python libraries and modules are imported here. This includes packages for data manipulation, visualization, model building, and evaluation.

3.Baseline Implementation of Logistic Regression

3.1 Algorithm Overview

Logistic Regression is a linear classifier that models the probability of the positive class using the sigmoid of a linear combination of features. We optimize the binary cross-entropy loss via gradient descent.

3.2 Utilities

Helper functions for loss calculation and other utilities used in the model.

Helper functions for loss calculation and other utilities used in the model.

Squared Error

The squared error between actual and predicted values:

- $(y - \hat{y})^2$

Mean Squared Error (MSE)

The average squared error across all samples:

- $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

Binary Cross-Entropy (BCE)

The main loss function for logistic regression, measuring the difference between true labels and predicted probabilities:

- $-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right]$

In code, we use a small constant `EPS` to avoid numerical issues with $\log$.

3.3 Base Classes

Implementation of base classes for estimators and regression models. BaseEstimator handles input validation and provides a standard interface for fit and predict methods.
BasicRegression extends BaseEstimator and implements the main logic for models trained with gradient descent, including regularisation and intercept handling.

3.4 Logistic Regression Class

Implementation of the custom Logistic Regression class, including the sigmoid function and prediction logic, for binary classification, using gradient descent to optimize the binary cross-entropy loss.

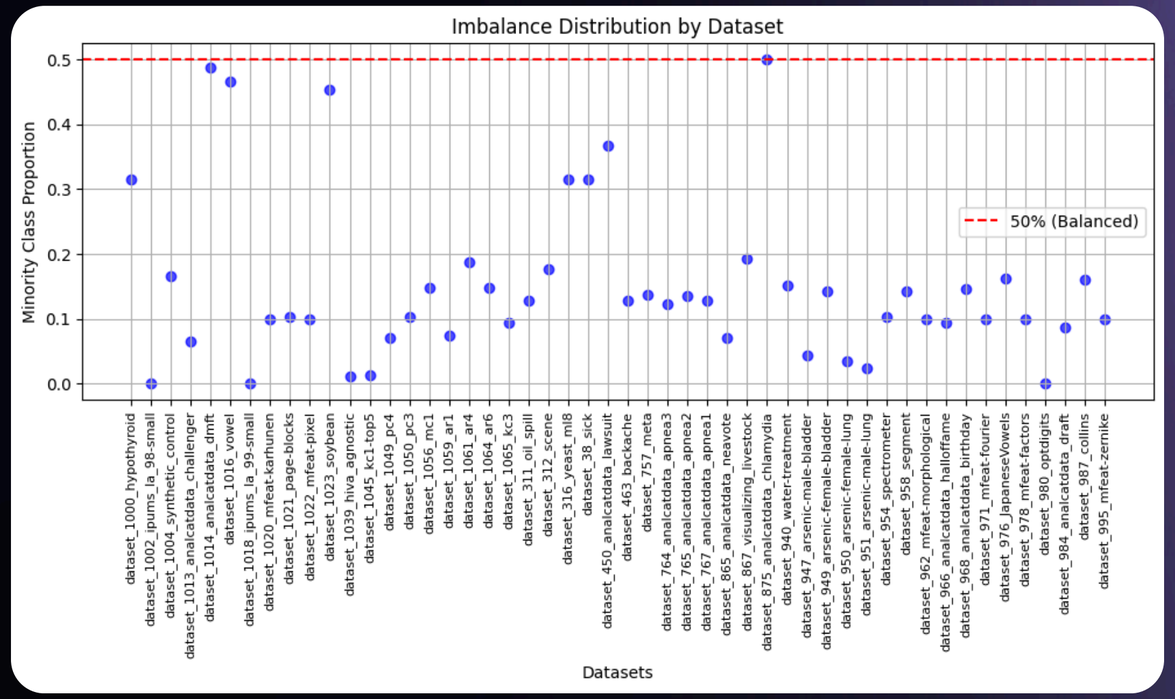
```python
class LogisticRegression(BasicRegression):
    """Binary logistic regression with gradient descent optimizer."""

    def init_cost(self):
        self.cost_func = binary_crossentropy

    def _loss(self, w):
        loss = self.cost_func(self.y, self.sigmoid(np.dot(self.X, w)))
        return self._add_penalty(loss, w)

    @staticmethod
    def sigmoid(x):
        return 0.5 * (np.tanh(0.5 * x) + 1)

    def _predict(self, X=None):
        X = self._add_intercept(X)
        return self.sigmoid(X.dot(self.theta))
```

4. Data Loading and Preprocessing

This section covers loading the datasets, identifying the binary target column, and preprocessing steps such as encoding categorical variables.

4.1 Discover all CSV files
4.2. Inspect shapes and target-column of each CSV
4.3 Data Preprocessing and Dataset Selection
4.3.1 Find binary target column (2 unique values) in a DataFrame
4.3.2 Minimal preprocessing:
- Try to convert object columns to numeric
- Apply get_dummies to X
- Apply LabelEncoder to y

4.3.3 Load all CSVs and keep only those with a binary target column
4.4 Imbalance Distribution by Dataset
4.5. Class Frequencies Across Datasets



4.4.1. Dataset Imbalance Analysis

Overview:
- X-axis: Dataset names
- Y-axis: Proportion of minority class

Key Insights:
- Severe Imbalance: Most datasets below 0.2 (strong imbalance).
- Few Balanced: Only a small number approach balance (0.5 line).
- Extreme Cases: Several datasets have proportions near zero (extreme imbalance).

Implication:
Highly imbalanced datasets challenge standard classifiers, emphasizing the importance of specialized imbalance-aware methods.

4.5.1 Class Frequencies Across Datasets (Stacked Bar Plot)

What it shows:

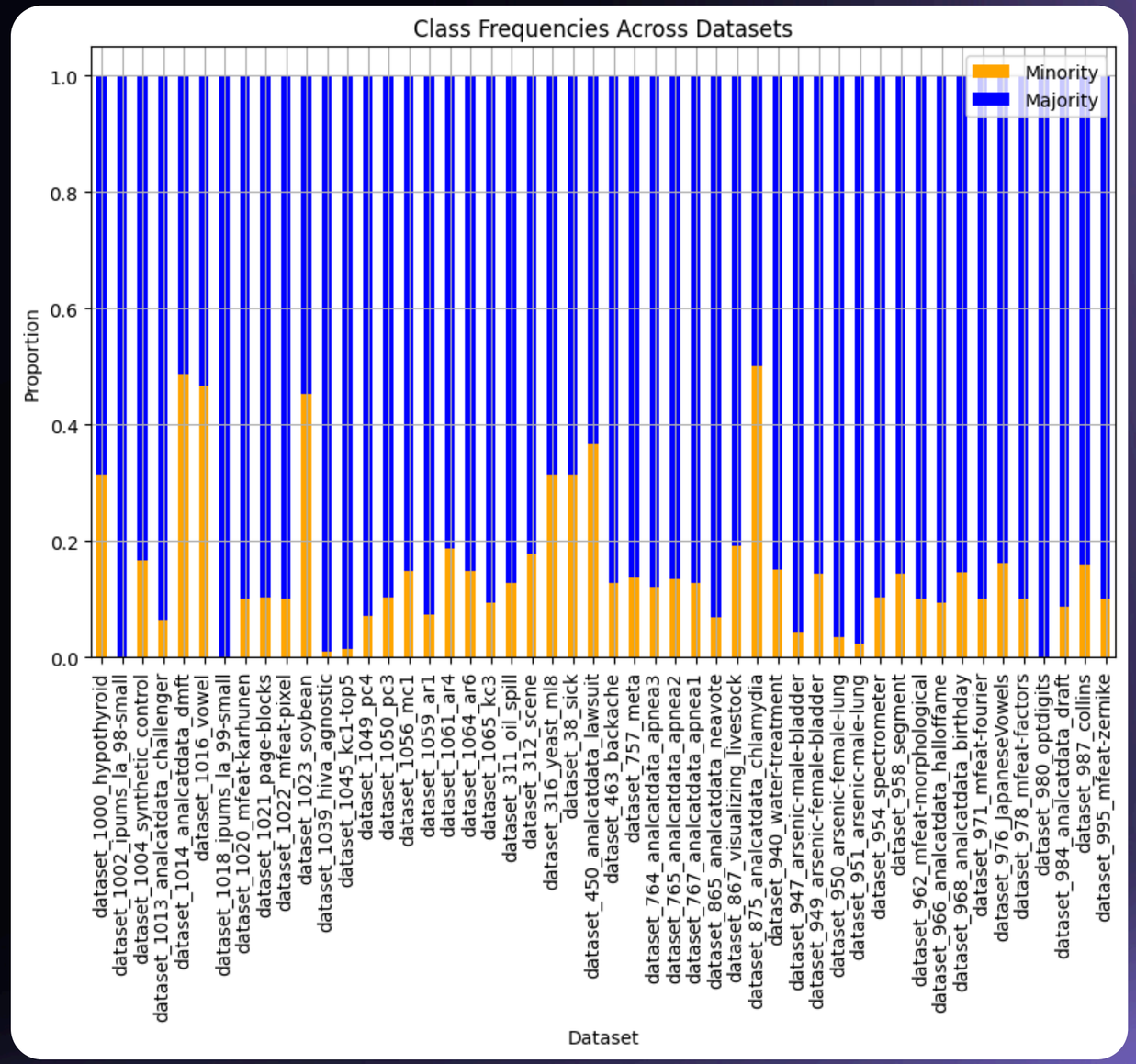
- Each bar represents a dataset, split into two colors:
  - Blue: Majority class
  - Orange: Minority class

Key Points:

- Majority Class Dominance: Most datasets are heavily dominated by the majority class (often 80–90%).
- Minority Class Scarcity: Minority class proportions are typically very small, confirming severe imbalance.
- Rare Balanced Cases: Few datasets have a balanced class distribution.

Implications:

- High imbalance can inflate accuracy by always predicting the majority class, misleading model evaluation.



5. Baseline Model Training and Evaluation

Here, the baseline model is trained and evaluated on each dataset. Performance metrics are computed and summarized.

5.1 Setup

5.2 Training and Evaluating the Baseline Model

Training the baseline model on each dataset and collecting evaluation metrics.

- Train/Test Split: For each dataset, we split the data into training (80%) and test (20%) sets.
- Model Training: Trained a custom Logistic Regression model (implemented from scratch) on the training data.
- Prediction: Used the trained model to predict probabilities and class labels on the test set.
- Comprehensive Evaluation: Calculated multiple performance metrics on the test set, including:
  - Accuracy, Precision, Recall, F1-score
  - ROC AUC, PR AUC (for probability ranking)
  - Matthews Correlation Coefficient (MCC)
  - Balanced Accuracy (BACC)
  - Confusion Matrix and Classification Report
- Results Aggregation:Stored all metrics for each dataset in a results table (DataFrame), and computed summary statistics (mean and standard deviation) across datasets.
- Visualization:Plotted boxplots to show the distribution of metrics across all datasets, highlighting performance variability and the challenges of imbalanced data.

5.2.1 Analysis of Baseline Results
5.3 Mean Confusion Matrix
5.3.1 Compute mean confusion matrix across all datasets
5.3.2 Visualize mean confusion matrix
5.3.3. Mean Confusion Matrix Analysis
5.4 Distribution of Confusion Matrix Elements
5.4.1 Extract confusion matrix elements
5.4.2 Visualize distribution of confusion matrix elements
5.4.3. Analysis: Distribution of Confusion Matrix Elements

5.5. Conclusion: Baseline Logistic Regression on Imbalanced Data

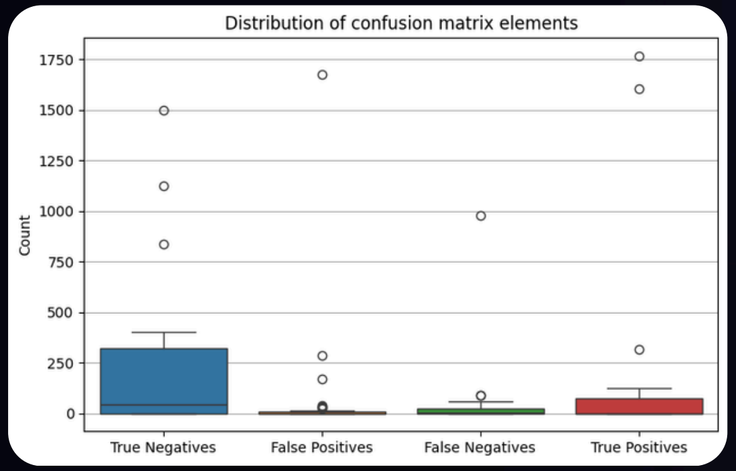
→ Accuracy is high (0.77) but misleading, as the model mostly predicts the majority class.

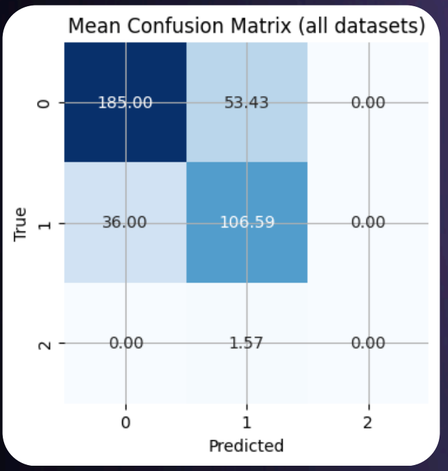
→ Precision (0.35), Recall (0.46), and F1 (0.36) are low and highly variable, indicating poor minority class detection.

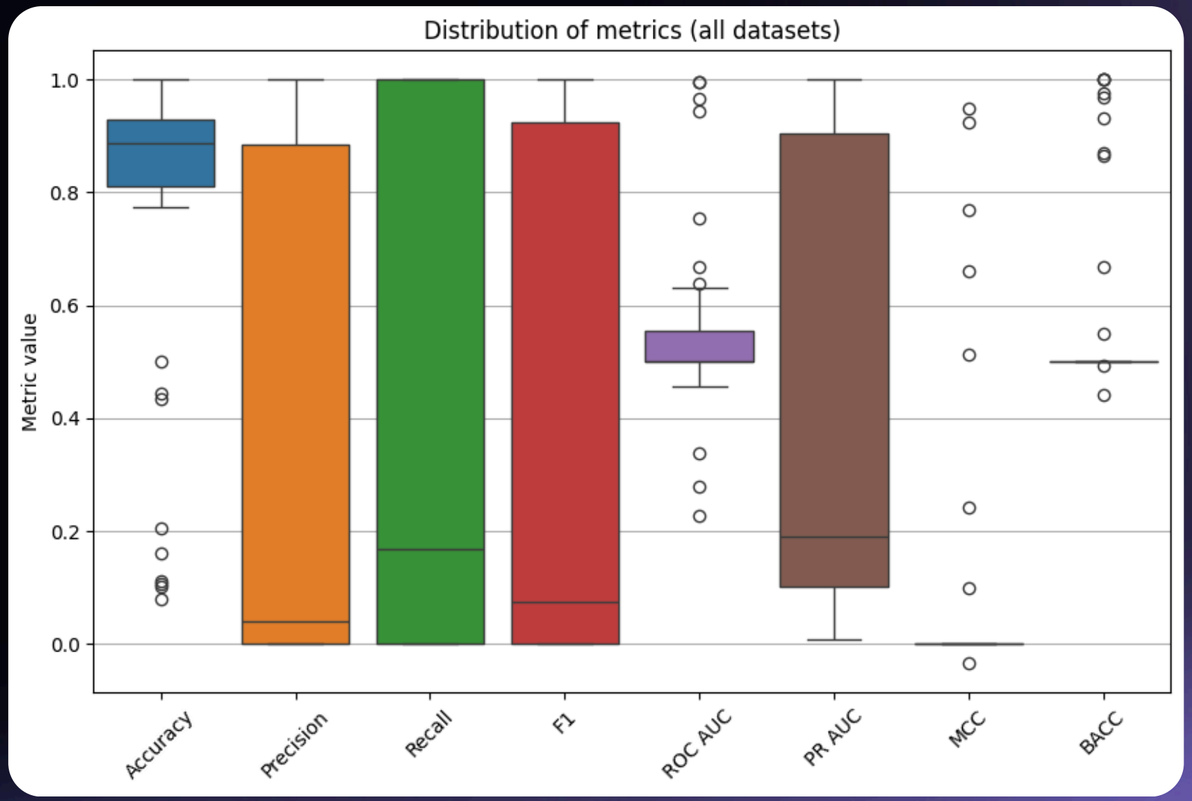
→ ROC AUC (0.55) and PR AUC (0.45) show near-random performance.

→ MCC (0.09) and Balanced Accuracy (0.59) confirm prediction imbalance.

→ The confusion matrix reveals frequent misclassification of minority samples as majority.







6. Focal Loss Variant: Implementation and Evaluation

6.1 LogisticRegressionFocal Class Definition

Focal Loss helps the model focus on hard-to-classify (often minority class) examples by reducing the loss contribution from easy samples.
The gamma parameter controls this effect: higher values put more focus on difficult cases. The implementation supports regularisation and uses gradient descent, similar to the baseline model.

6.2 Gamma Tuning and Evaluation

6.2.1 Function for gamma tuning via 3-fold CV

6.2.2 Evaluate Focal Loss on each dataset using the same split as baseline

	Mean	Std
Accuracy	0.763492	0.262597
Precision	0.494779	0.210603
Recall	0.532615	0.128328
F1	0.484066	0.182588
ROC AUC	0.556994	0.162683
PR AUC	0.446449	0.355466
MCC	0.072329	0.192374
BACC	0.543587	0.143685

Dataset	γ	Accuracy	Precision	Recall	F1	ROC AUC	PR AUC	MCC	BACC
ıtaset_1002_ipums_la_98-small	1.0	0.742819	0.499551	0.371658	0.426217	0.371658	0.999161	-0.015186	0.371658
dataset_1004_synthetic_control	0.0	0.833333	0.416667	0.500000	0.454545	0.485000	0.168333	0.000000	0.500000
et_1013_analcatdata_challenger	0.0	0.928571	0.464286	0.500000	0.481481	0.500000	0.071429	0.000000	0.500000
dataset_1014_analcatdata_dmft	0.0	0.475000	0.476590	0.479049	0.462916	0.484991	0.507406	-0.044293	0.479049
dataset_1016_vowel	0.0	0.702020	0.700495	0.699446	0.699807	0.735029	0.659197	0.399940	0.699446
taset_1018_ipums_la_99-small	1.0	1.000000	1.000000	1.000000	1.000000	NaN	NaN	0.000000	1.000000

6.3.1.1 Analysis Metrics Table per Dataset and Statistical Summary (Mean and St)

→ Gamma needs tuning per dataset; no universal best value.
→ Accuracy and F1 vary widely — some datasets improve, others don't.
→ ROC AUC and PR AUC often low or NaN → minority class still hard to detect.
→ MCC and BACC are weak → class balance remains a challenge.
→ High variability across metrics → inconsistent results.
→ Conclusion: Focal Loss helps in some cases, but is not a universal fix for imbalance.

6.3.2.1 Analysis: Side-by-side comparison table

- Focal Loss provides a clear benefit for most imbalanced datasets, especially where the baseline model fails to predict the minority class.
- Gamma tuning is important: higher values help in some cases, but not all.
- Occasional performance drops suggest the need for careful validation and possibly combining Focal Loss with other imbalance strategies.

	F1		ROC AUC		γ
Model	Baseline	FocalLoss	Baseline	FocalLoss	FocalLoss
Dataset					
dataset_1002_ipums_la_98-small	1.000000	0.426217	NaN	0.371658	1.0
dataset_1004_synthetic_control	0.000000	0.454545	0.500000	0.485000	0.0
dataset_1013_analcatdata_challenger	0.000000	0.481481	NaN	0.500000	0.0
dataset_1014_analcatdata_dmft	0.000000	0.462916	0.550957	0.484991	0.0
dataset_1016_vowel	0.975124	0.699807	0.995612	0.735029	0.0
dataset_1018_ipums_la_99-small	0.999717	1.000000	0.500000	NaN	1.0
dataset_1020_mfeat-karhunen	0.931507	0.771429	0.995803	0.915417	0.0
dataset_1021_page-blocks	0.000000	0.788637	0.500000	0.676905	0.0
dataset_1022_mfeat-pixel	0.000000	0.842030	0.500000	0.966944	0.0
dataset_1023_soybean	0.209637	0.072896	NaN	NaN	1.0

7. Comparative Analysis: Metrics and Visualizations

Comparison of the baseline and Focal Loss models using various metrics and visualisations such as boxplots, heatmaps, and radar plots.

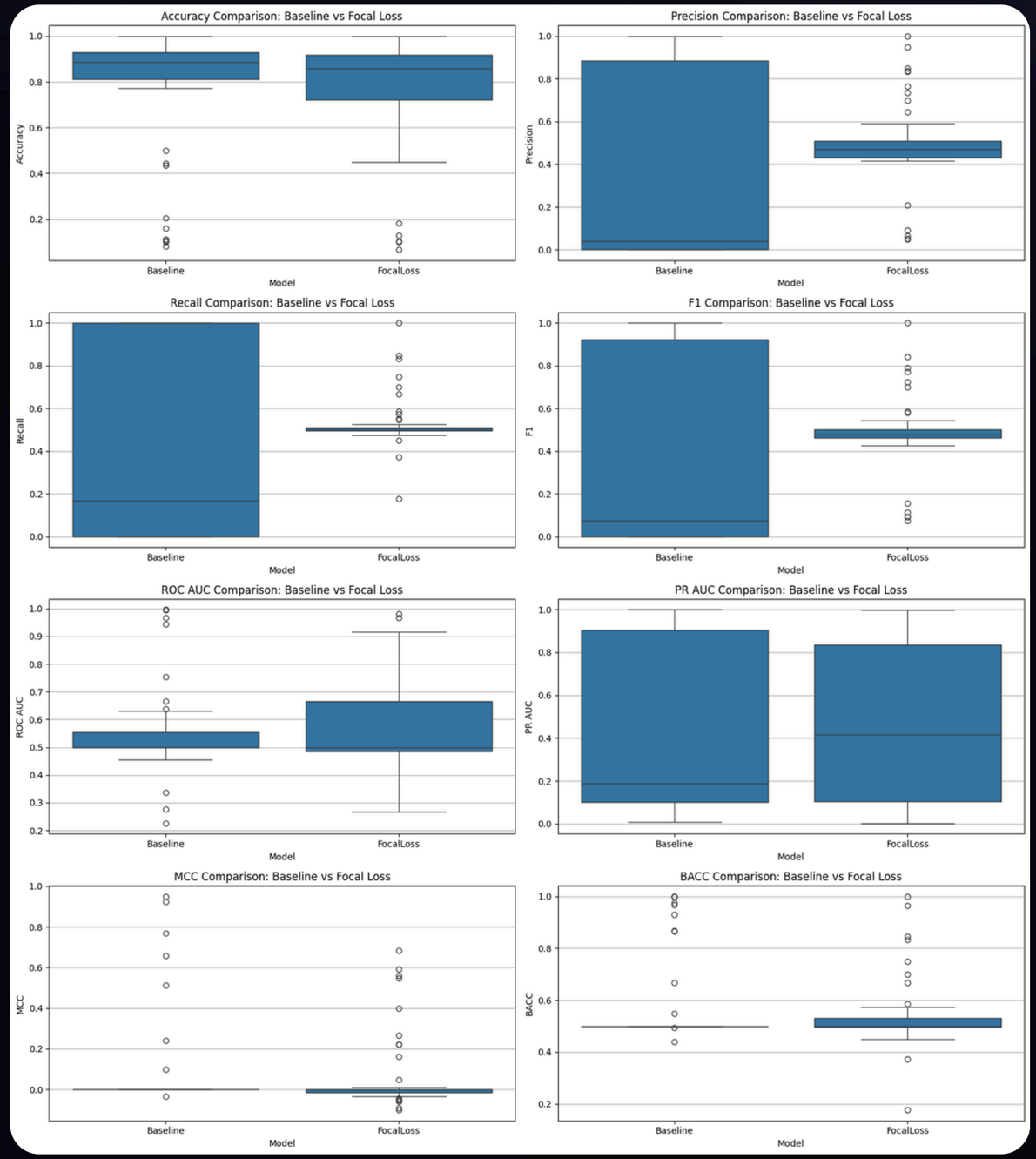
7.1 Distribution of Metrics

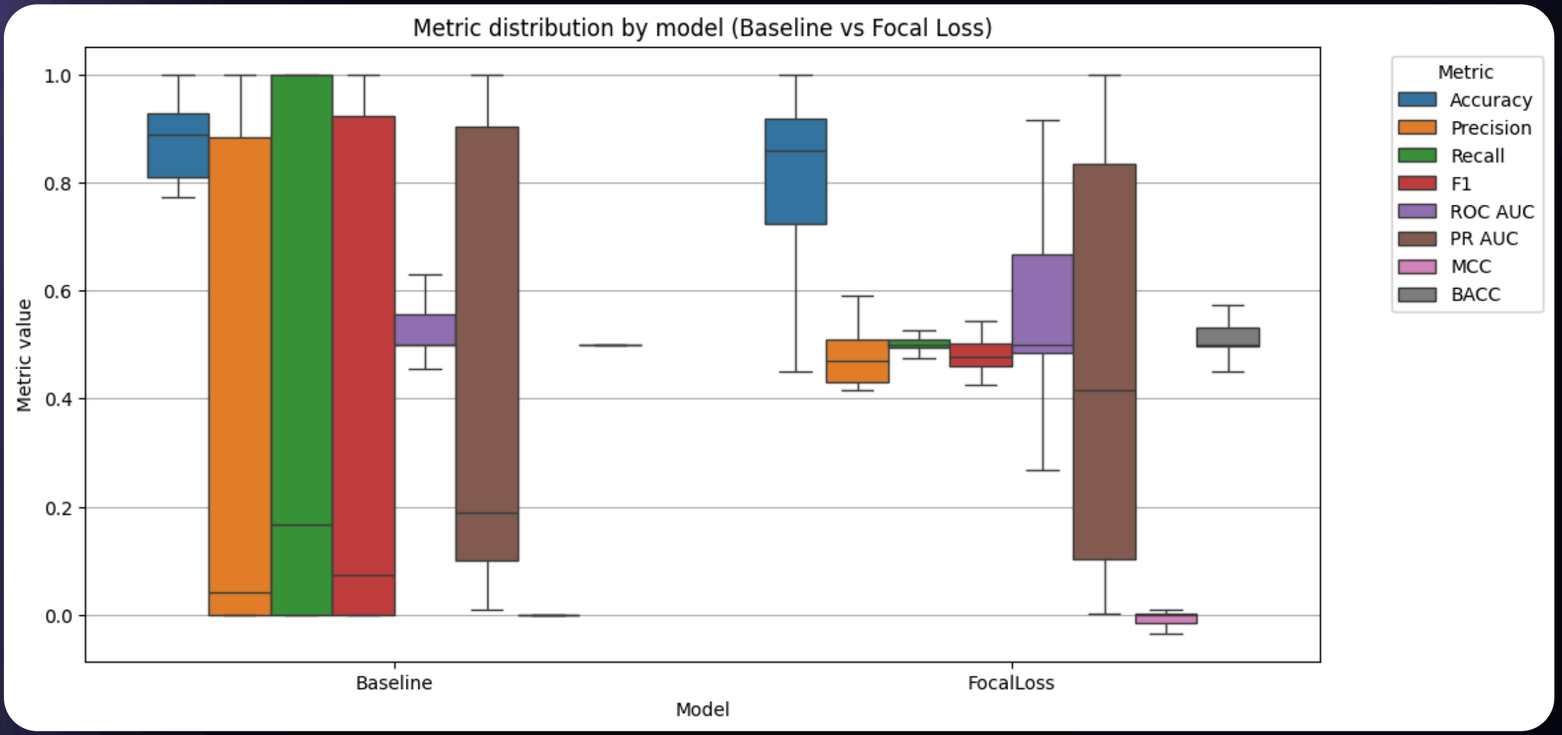
This subsection presents boxplots comparing the distribution of key metrics (Accuracy, Precision, Recall, F1, ROC AUC, PR AUC, MCC, BACC) for both models across all datasets.

7.1.1. Each metric

7.1.2 All Metrics

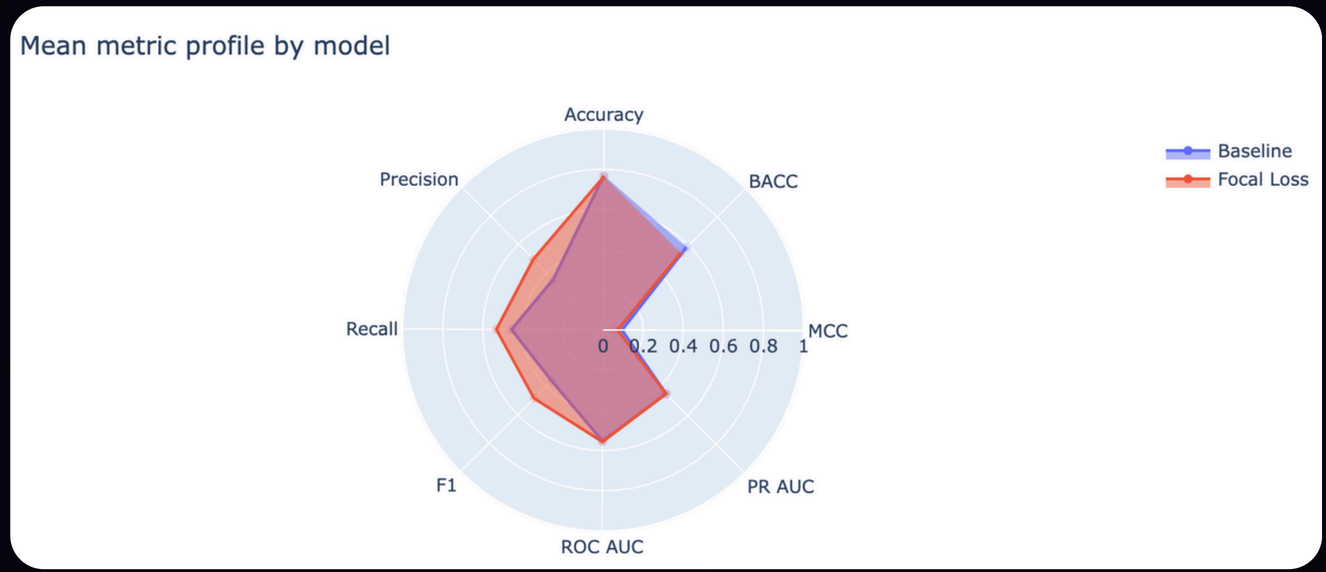
7.1.3 Radar Plot of Mean Metrics



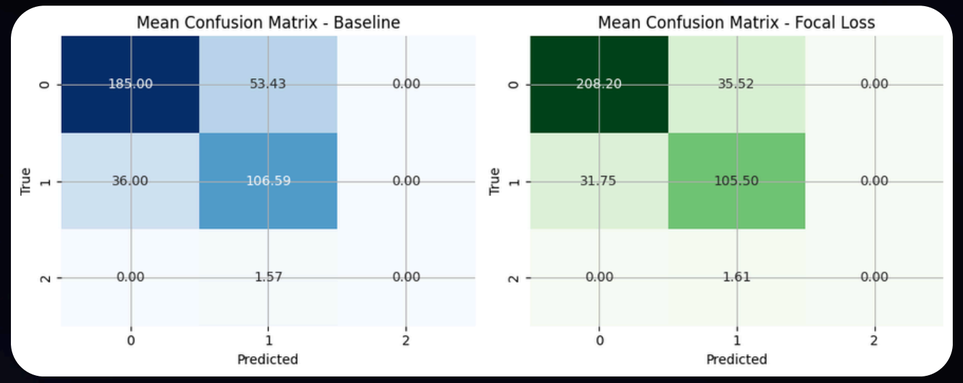


7.1.4 Analysis - Metric distribution by model

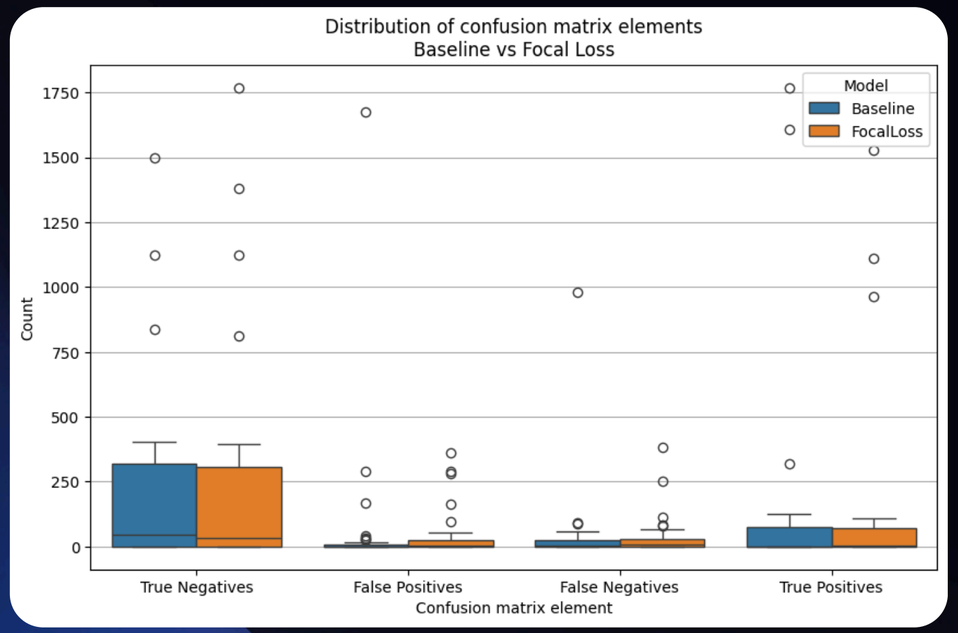
- **Focal Loss** provides substantial and consistent improvements in metrics that matter for imbalanced classification (Recall, F1, PR AUC, MCC).
- **Baseline** model's high accuracy is misleading; it fails to capture minority class performance.
- **Variability** is reduced with Focal Loss, indicating more robust and reliable performance across diverse datasets.
- **Trade-off:** Focal Loss may slightly reduce Precision or Accuracy, but this is expected and acceptable when the goal is to improve minority class detection.



7.2 Mean Confusion Matrix Comparison



7.3 Distribution of Confusion Matrix Elements

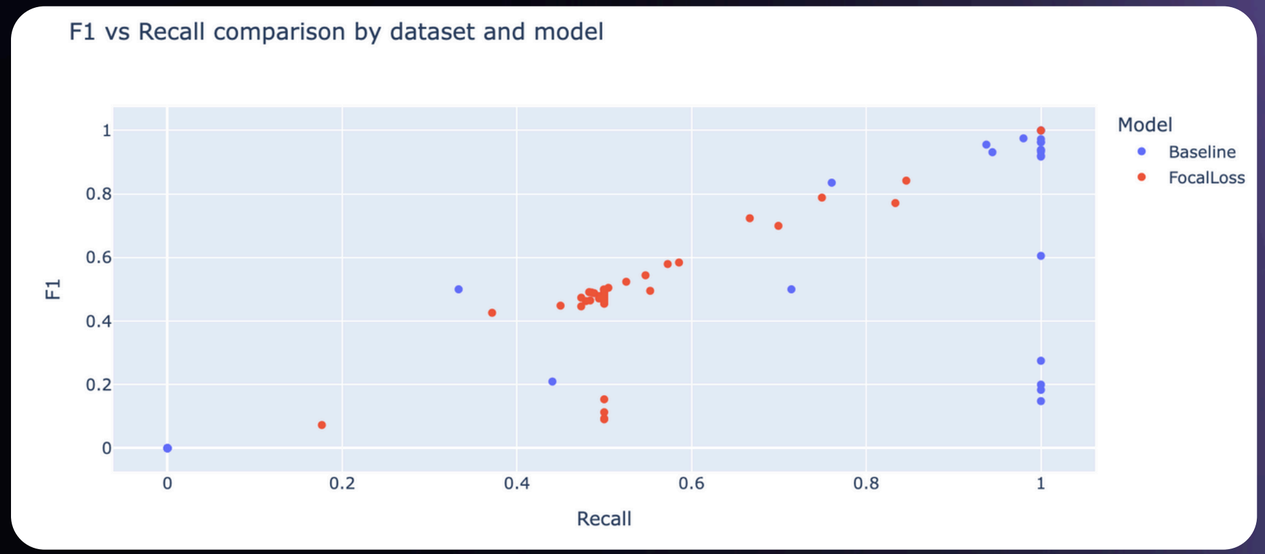


7.4 Analysis of Mean Confusion Matrices: Baseline vs Focal Loss

- TN: Both models have high TN; Focal Loss shows less variability, indicating more consistency.
- FP: Focal Loss has lower median FP and narrower spread, meaning fewer false alarms and better precision.
- FN: Slightly higher in Focal Loss, showing a trade-off, but still more stable than Baseline.
- TP: Focal Loss achieves higher and more consistent TP, improving minority class detection.
- Baseline: High FP (53.43) and FN (36.00) → frequent misclassifications.
- Focal Loss: Higher TN (208.20), lower FP (35.52), slightly lower FN (31.75) → better class separation.
- More diagonal confusion matrix in Focal Loss → stronger overall classification.
- Conclusion: Focal Loss achieves better balance between precision and recall, reduces errors, and is more robust and reliable for imbalanced datasets.

7.5 F1 vs Recall Scatter Plot
7.5.1. Analysis: F1 vs Recall Comparison by Dataset and Model

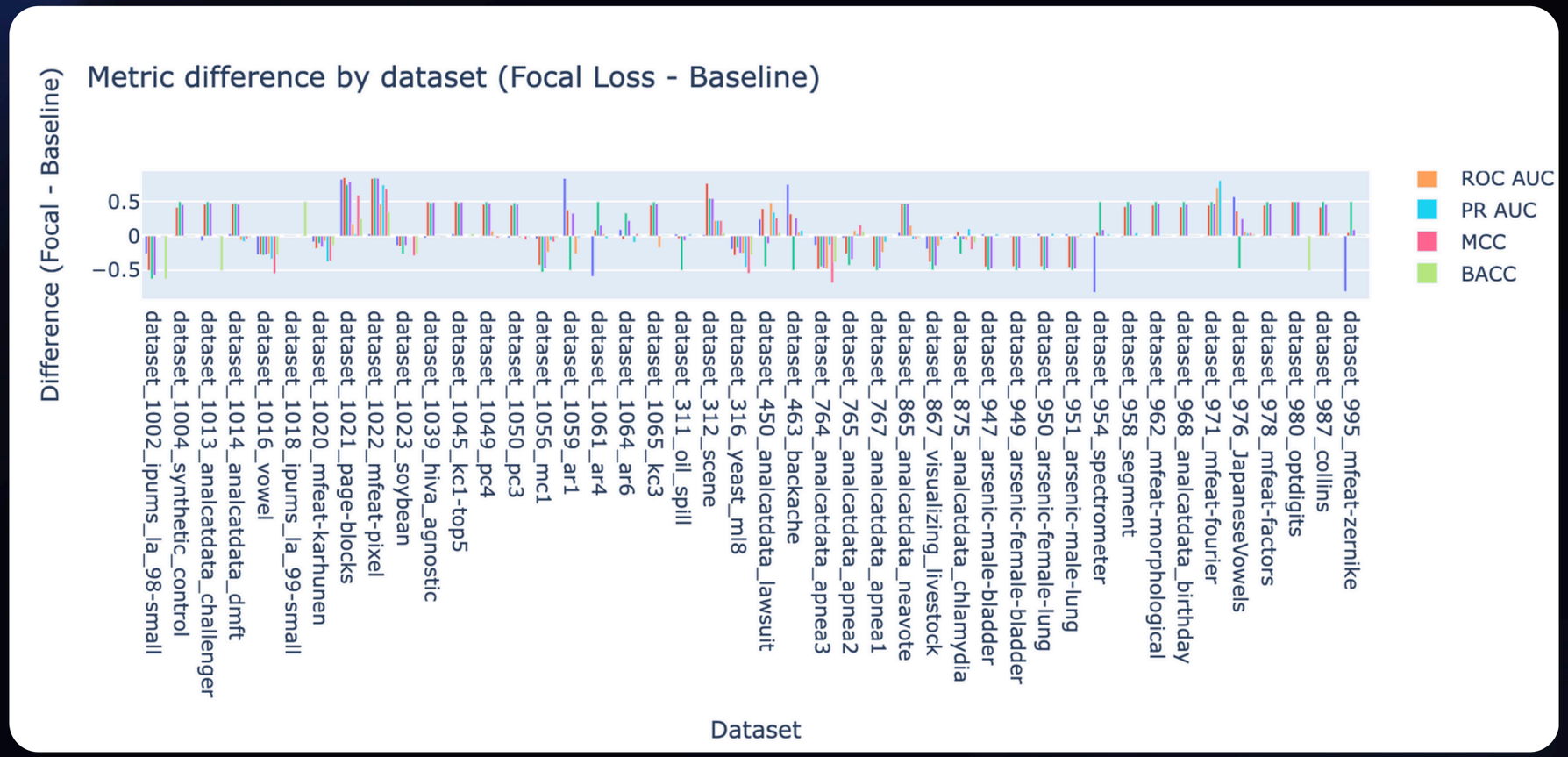
- Baseline Model: Many datasets show Recall and F1 near zero → fails to detect the minority class.
- Focal Loss Model: Most datasets have Recall > 0.4 and higher F1 → better minority class detection.
- Clusters: Focal Loss clusters in high Recall/F1 region; Baseline clusters in low performance zone.
- Trade-off: Focal Loss may reduce Precision to gain Recall → but results in better overall balance.
- Summary: Focal Loss consistently outperforms the Baseline in Recall and F1, showing greater robustness for imbalanced datasets.

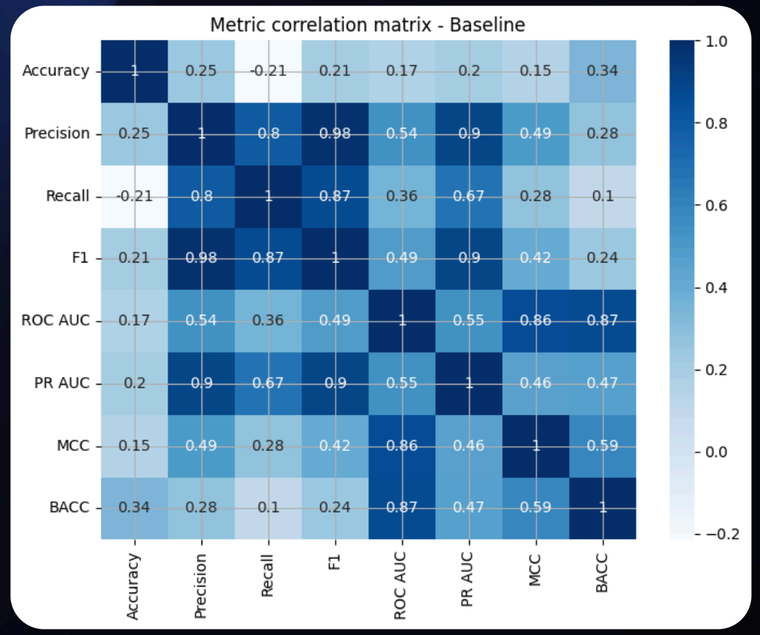


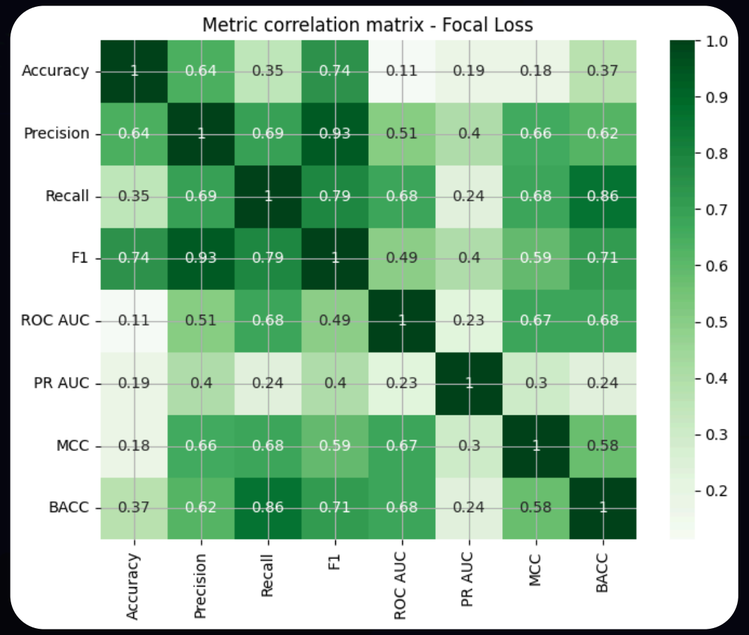
7.6 Metric Difference by Dataset

7.6.1 Analysis: Metric difference by dataset

Focal Loss usually increases Recall and sometimes Precision/Accuracy, but results depend on the dataset. Gains are not guaranteed for all cases.







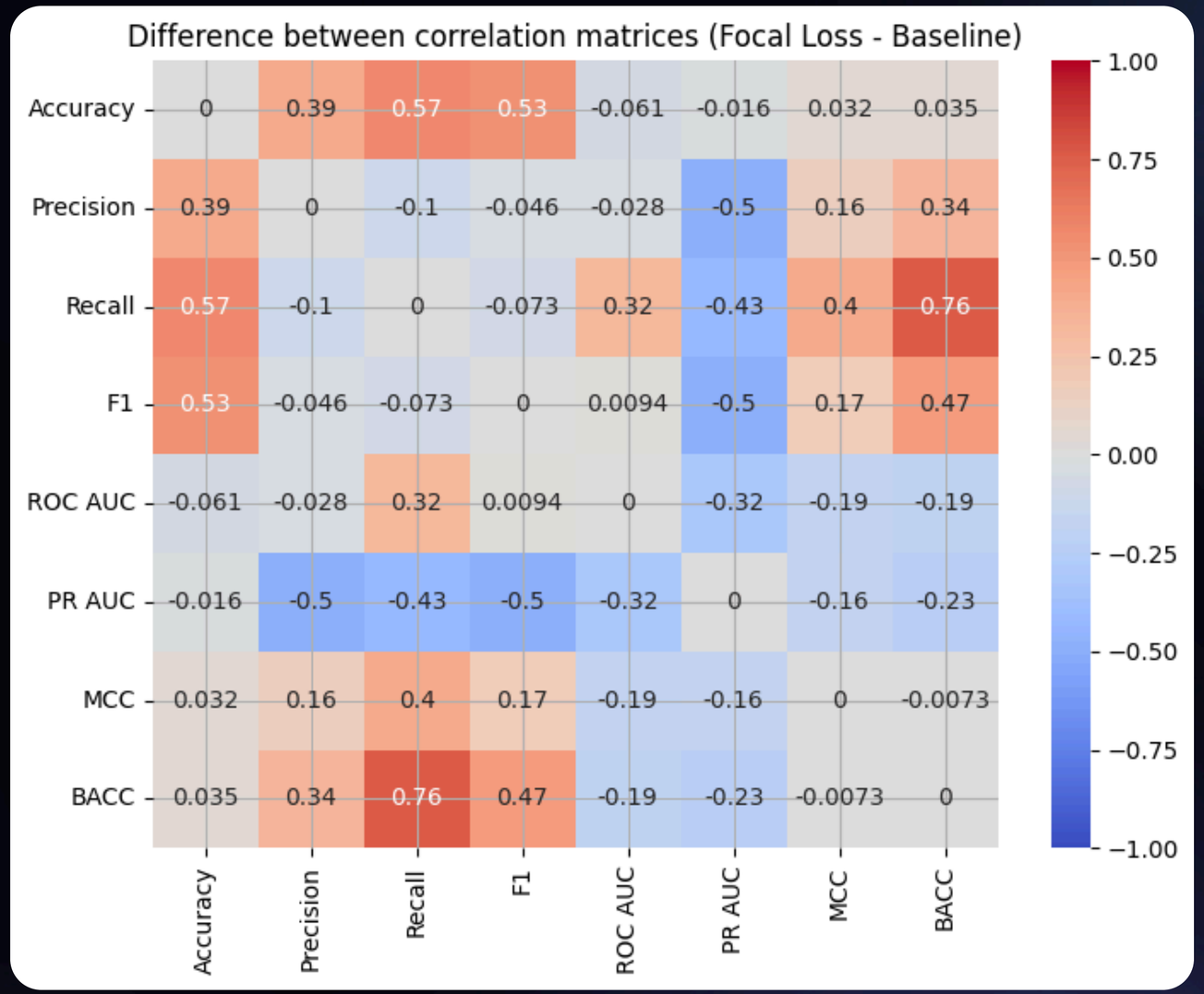
7.7. Correlation Matrices of Metrics
7.7.1 Analysis: Correlation Matrices of Metrics

- Focal Loss: Correlations between key metrics (Recall, F1, BACC) are higher and more balanced, showing metrics move together more consistently. Accuracy is better aligned with Recall and F1.
- Baseline: Correlations are weaker and more scattered, especially for Accuracy vs Recall (even negative), indicating metrics often disagree, typical in imbalanced settings.
- Summary: Focal Loss produces more consistent and reliable metric relationships, reflecting more robust and fair model behavior across datasets.

7.8 Difference Between Correlation metrics.

7.8.1 Analysis: Difference Between Correlation Matrices

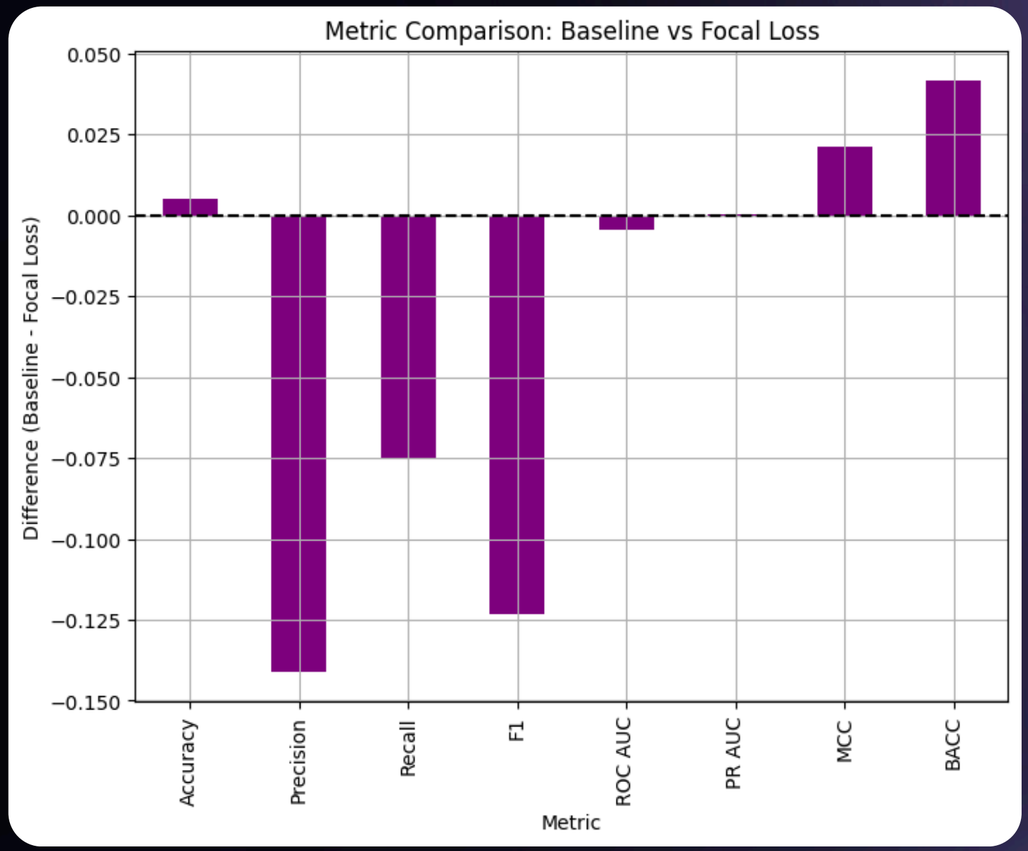
- What changed:
  - Focal Loss increases the correlation between Accuracy, Recall, F1, and BACC (more red in those cells), meaning these metrics move together more under Focal Loss.
- Where it dropped:
  - Correlations between Precision/PR AUC and other metrics decrease (more blue), showing less alignment with the rest.
- Summary:
  - Focal Loss makes key metrics (Recall, F1, BACC) more consistent with each other, but reduces the dependence of Precision/PR AUC on other metrics. This reflects a shift toward more reliable minority class detection.

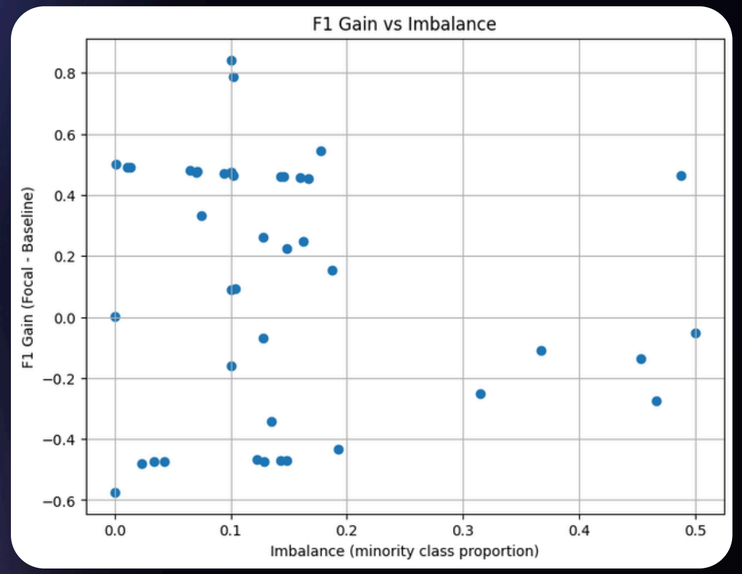


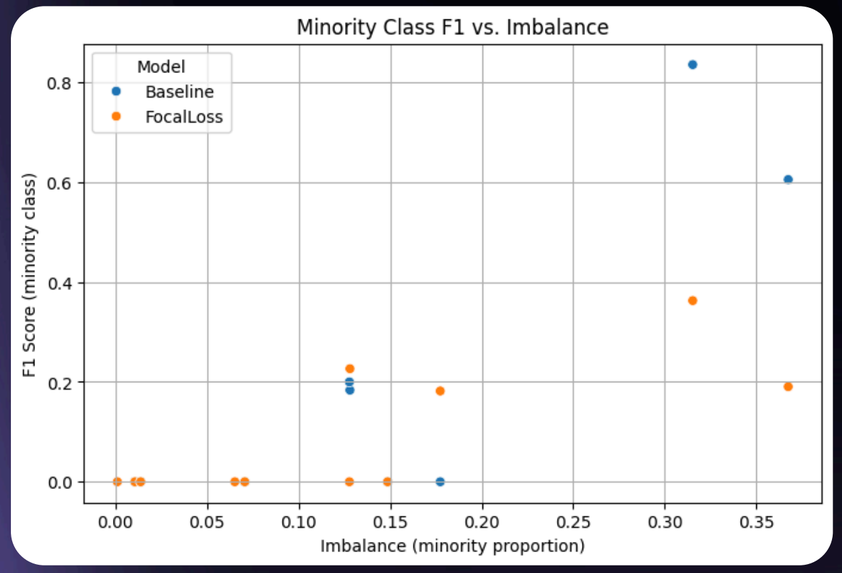
7.9 Metric Comparison for Baseline and Focal Loss models

7.9.1 Analysis: Metric Comparison for Baseline and Focal Loss models

- Recall, F1, BACC, MCC: Focal Loss outperforms Baseline (negative bars = Focal Loss higher).
- Precision: Baseline is higher, indicating Focal Loss trades some precision for recall.
- Accuracy, ROC AUC, PR AUC: Differences are minimal.
- Summary: Focal Loss improves recall, F1, and balanced metrics, making it better for imbalanced data, at the cost of slightly lower precision.







8. Advanced Visualizations and Statistical Analysis

8.1 Most Impacted Datasets (F1 Gain/Loss)

8.2 Correlation Between Imbalance and Focal Loss Gain

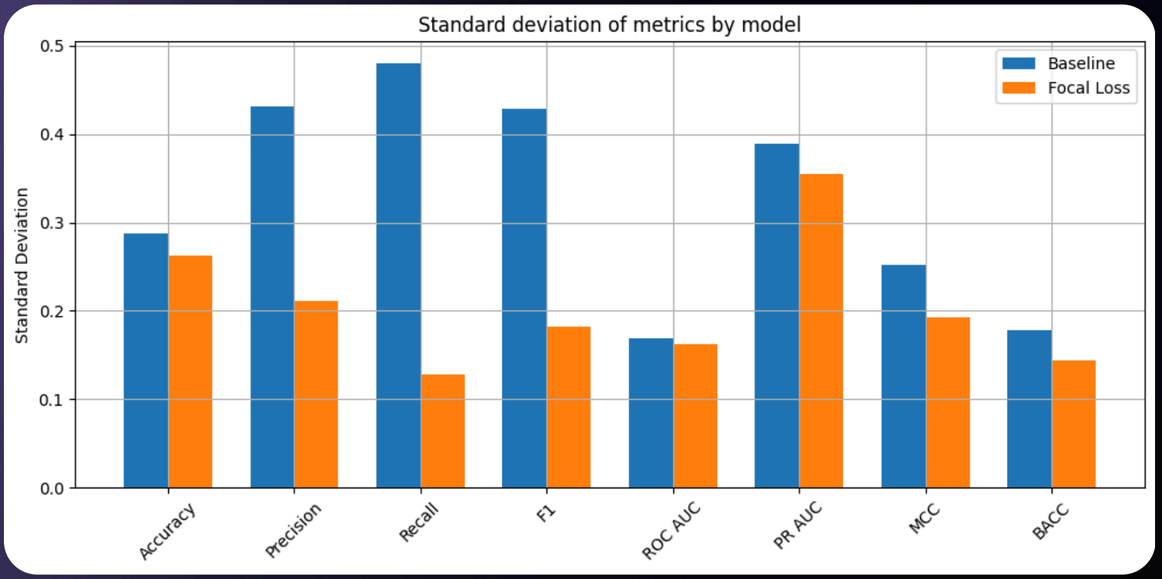
8.2.1 Analysis: F1 Gain vs. Imbalance:

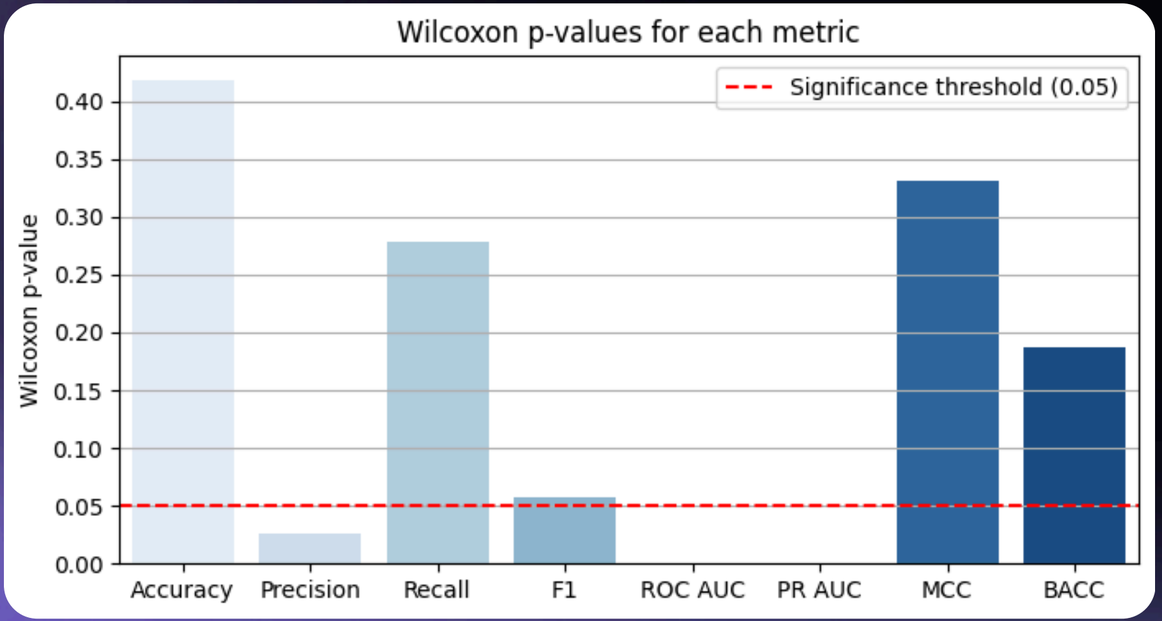
No strong correlation, Focal Loss can help or hurt regardless of imbalance level. Some highly imbalanced datasets see large F1 gains, others do not.

8.3. Impact of Imbalance on Minority Class Performance (F1, Recall, Precision)

8.3.1 Analysis: Minority Class F1 vs. Imbalance

Both models struggle as imbalance increases, but Focal Loss (orange) sometimes achieves higher minority F1. However, gains are not consistent for all datasets.





8.4 Robustness Analysis (Standard Deviation of Metrics)

8.4.1 Standard deviation of metrics by model:

Focal Loss consistently reduces metric variability across datasets, indicating more robust and stable performance than Baseline.

8.5 Statistical Significance Test (Wilcoxon Signed-Rank)

8.5.1 Visual Comparison of Wilcoxon P-Values

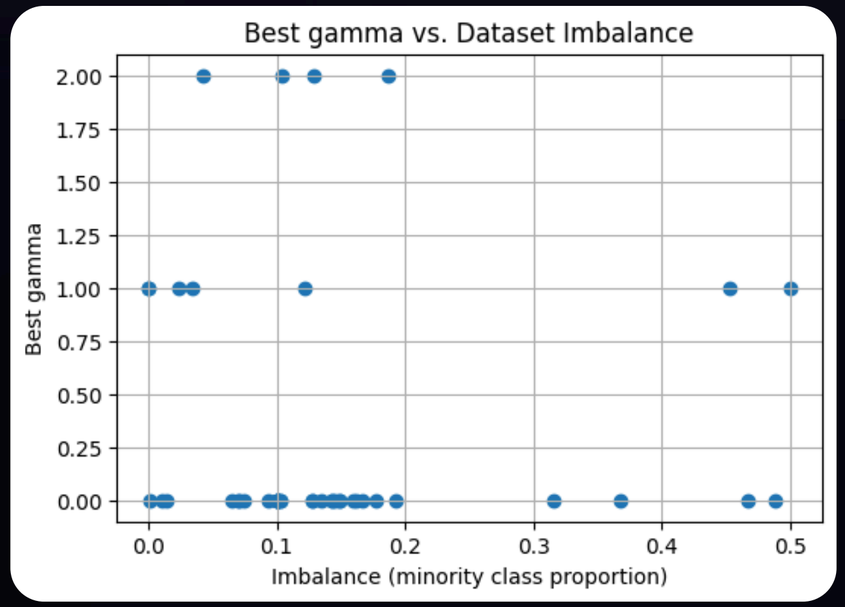
A bar plot of Wilcoxon p-values for each metric, with a threshold line at 0.05 to highlight statistically significant differences.

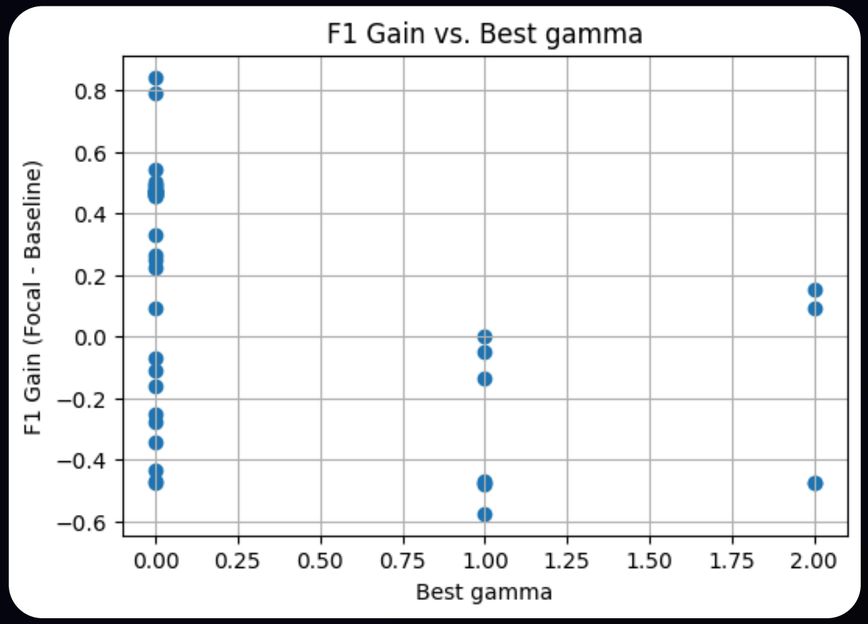
8.5.1.1 Analysis: Wilcoxon p-values for each metric:

Only Precision and F1 show statistically significant differences ($p < 0.05$) between Baseline and Focal Loss. Other metrics show no significant difference.

8.6 Effect Size Analysis (Rank-Biserial Correlation)
8.7 Gamma Value
8.7.1. Analysis: Gamma Value

1. Best gamma vs. Dataset Imbalance: No clear trend, optimal gamma values are spread across all imbalance levels. Focal Loss does not always require higher gamma for more imbalanced datasets.
2. F1 Gain vs. Best gamma: Most F1 gains occur at gamma=0 (i.e., no Focal Loss), but some datasets benefit from gamma=1 or 2. High gamma does not guarantee higher F1 gain.



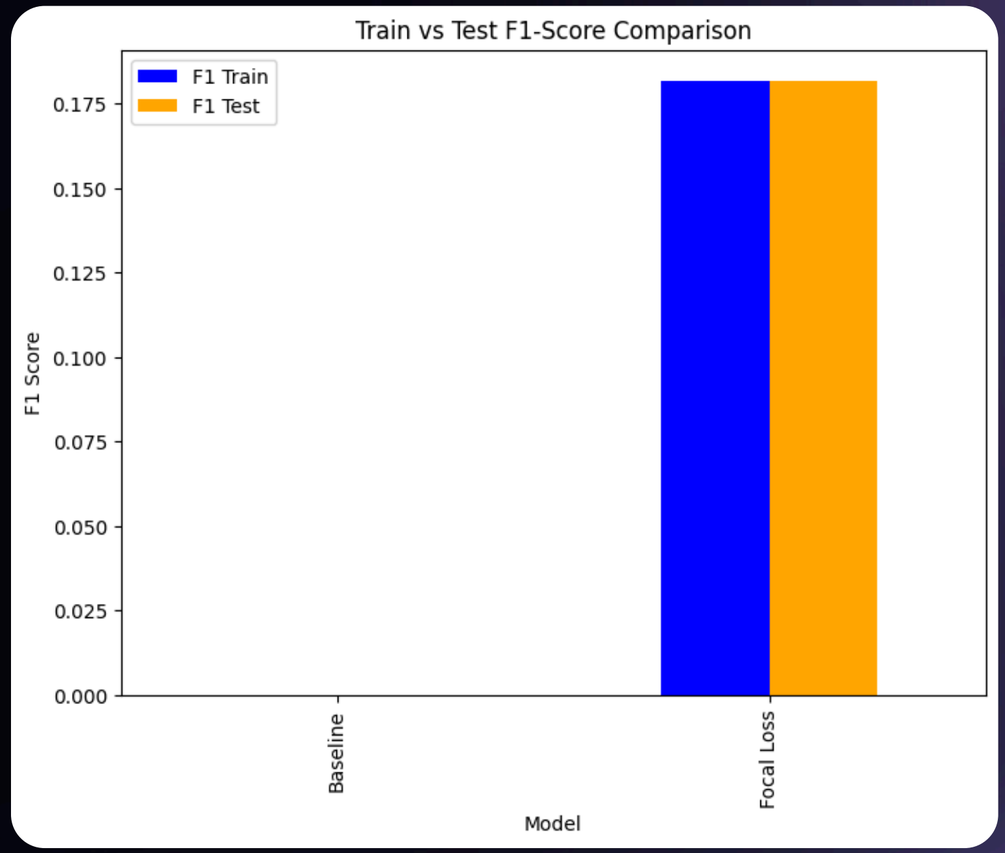


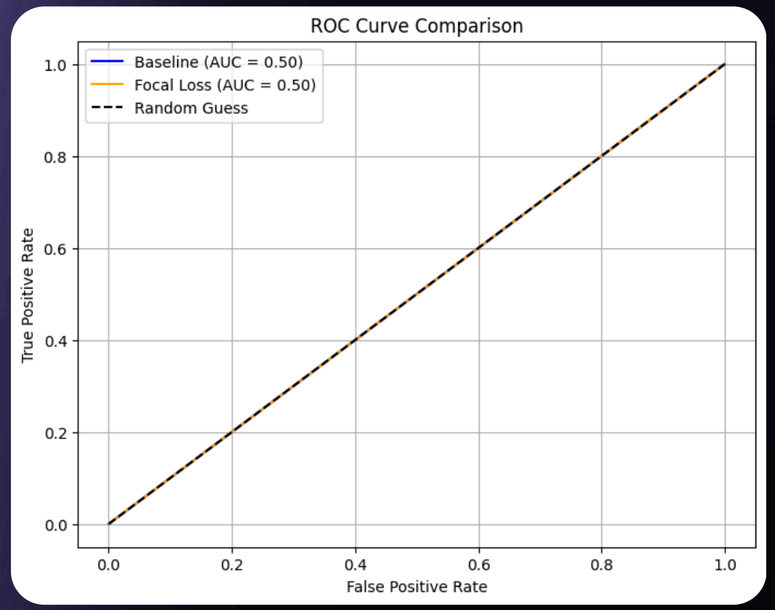
9.1 Overfitting Analysis (Train vs Test Performance)

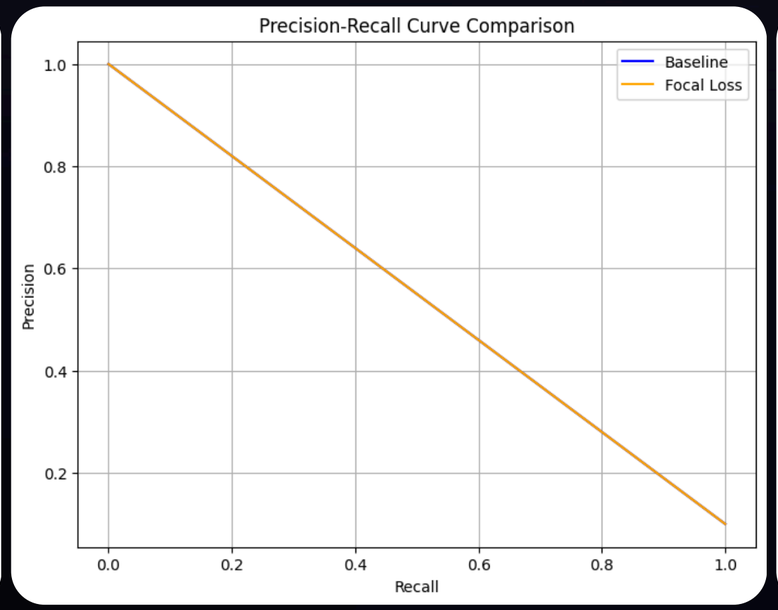
9.1.4. Comparison DataFrame

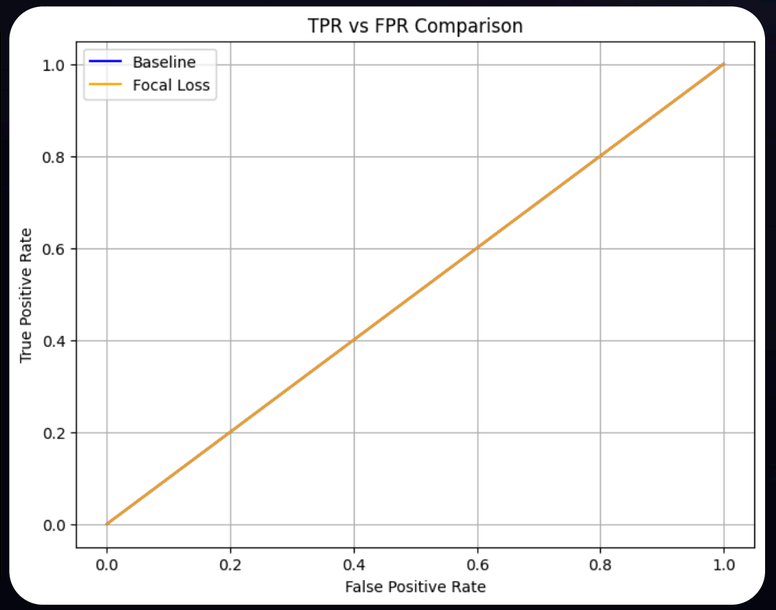
9.1.5 Analysis: Train vs Test F1-Score Comparison

- Both Baseline and Focal Loss models have very low F1-scores, indicating poor performance on this dataset.
- The F1-score for train and test is nearly identical for each model, showing no overfitting (no gap between train and test).
- Focal Loss does not show a clear advantage over Baseline in this case, both perform similarly and poorly.
- Conclusion: Both models struggle equally, likely due to severe class imbalance or dataset difficulty. No overfitting is observed.









10. Conclusions and References

10.1. ROC Curve for Both Models to show the trade-off between the true positive rate (TPR) and false positive rate (FPR) for different thresholds.

10.1.1 Analysis: ROC Curve Comparison

Both models have ROC curves on the diagonal (AUC ≈ 0.5), confirming no predictive power on this dataset.

10.2 Precision-Recall (PR) Curve for Both Models

10.2.2 Analysis: Precision-Recall Curve

Curves for both models overlap and are close to the diagonal, showing poor precision-recall trade-off (random-like)

10.3. True Positive Rate (TPR) vs False Positive Rate (FPR) for both models

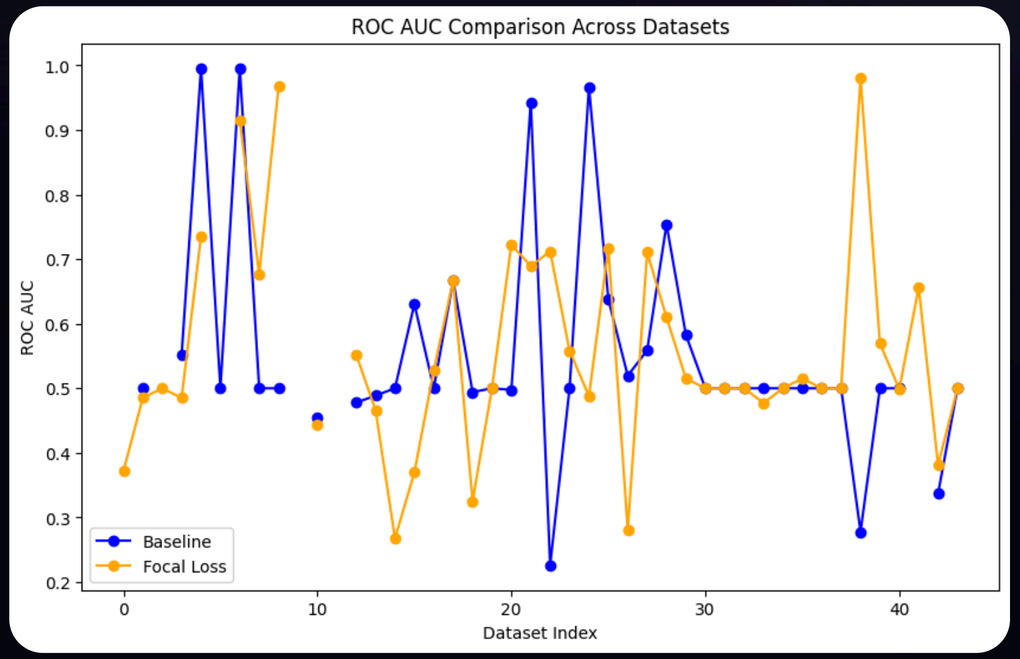
10.3.1 Analysis: TPR vs FPR

Both models overlap and follow the diagonal, indicating random performance (no discrimination).

10.4. ROC AUC Comparison Across Datasets for both models across all datasets to highlight performance consistency

10.4.1. Analysis: ROC AUC Comparison Across Datasets:

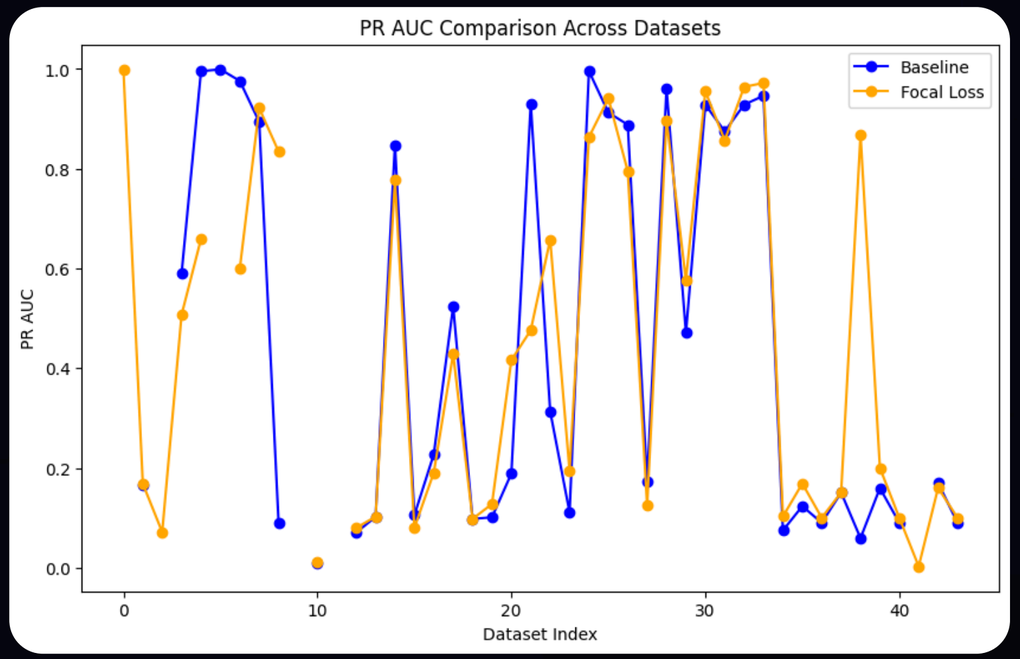
- ROC AUC fluctuates for both models, with neither model consistently superior.
- Focal Loss occasionally achieves higher ROC AUC, but Baseline is better in some datasets.

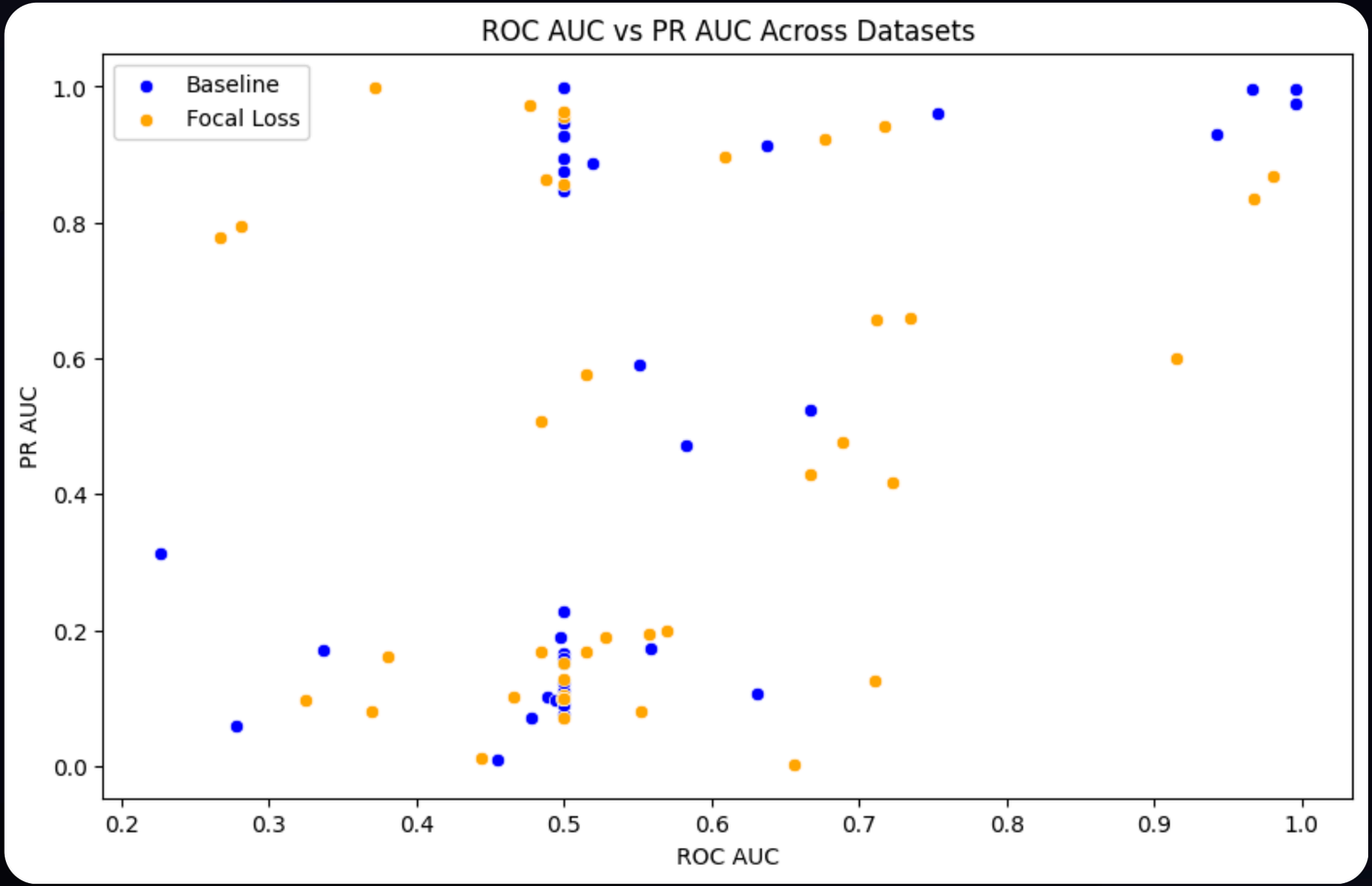


10.5 Precision-Recall AUC Comparison Across Datasets

10.5.1 Analysis: PR AUC Comparison Across Datasets

- PR AUC varies a lot by dataset for both models.
- Focal Loss (orange) sometimes outperforms Baseline (blue), but not consistently; lines cross frequently.

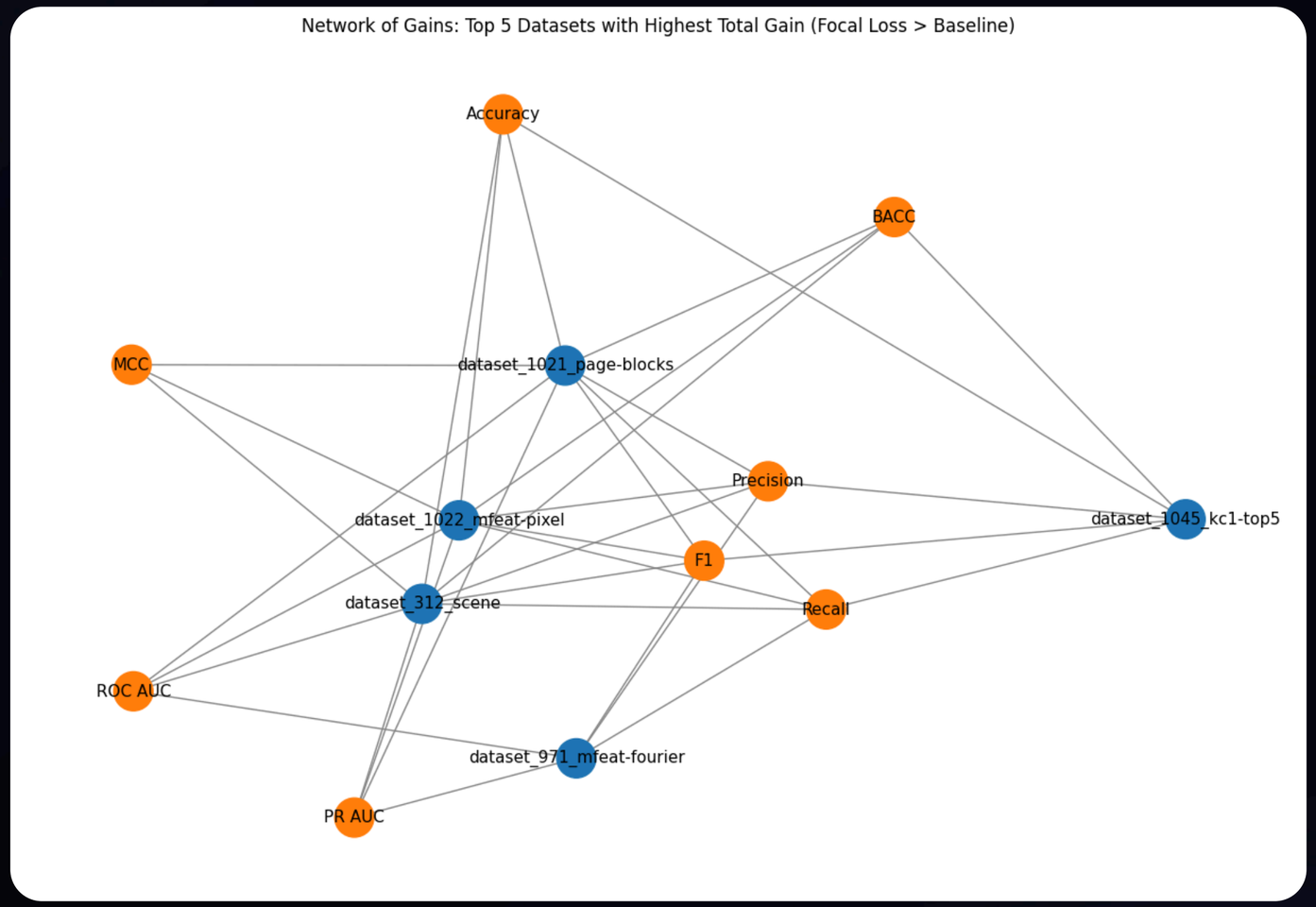




10.6 Dataset-Level Summary: ROC AUC and PR AUC

10.6.1 Analysis: ROC AUC vs PR AUC Scatter

- Both Baseline (blue) and Focal Loss (orange) models show wide spread in ROC AUC and PR AUC across datasets.
- No model dominates across all datasets; Focal Loss sometimes improves PR AUC, but not always.
- Summary: Neither Baseline nor Focal Loss model shows consistent or significant improvement in ROC AUC or PR AUC across all datasets. On some datasets, Focal Loss helps, but overall both models often perform close to random guessing.



10.7. Network of Gains: Top 5 Datasets with Highest Total Gain (Focal Loss > Baseline)

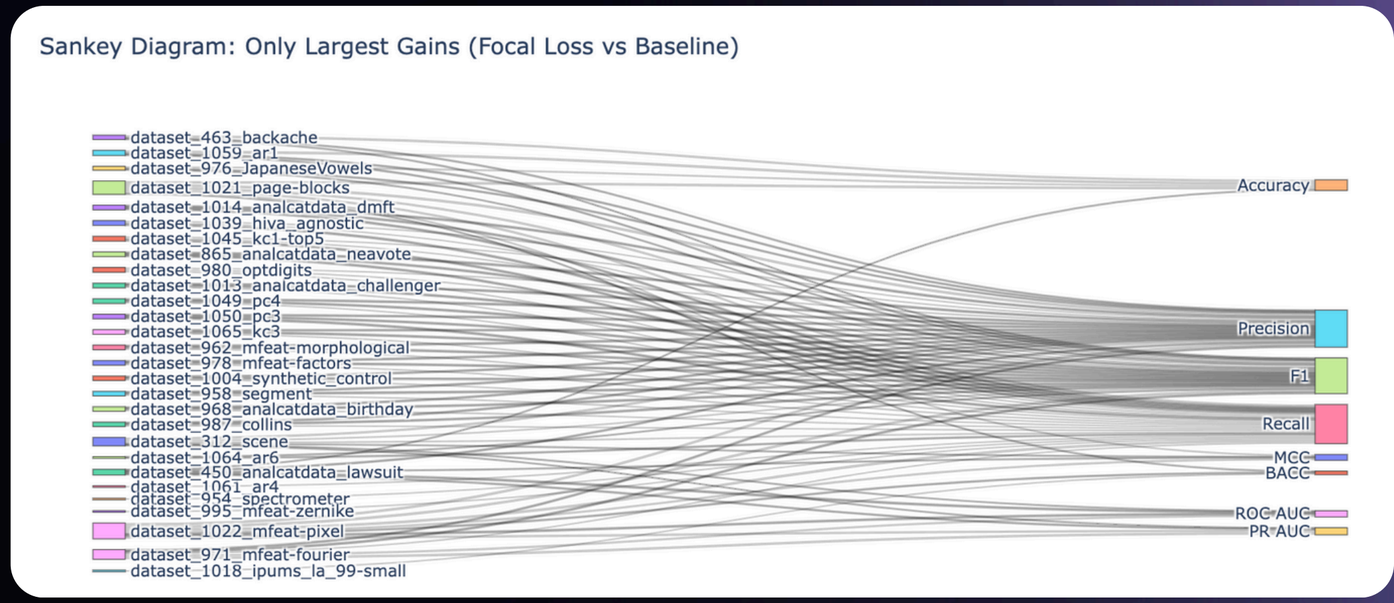
10.7.1 Insights: Network of Gains: Top 5 Datasets with Highest Total Gain (Focal Loss > Baseline)

- These top datasets benefited from Focal Loss across multiple metrics, not just one.
- Some datasets are connected to nearly all metrics, indicating broad improvement.

10.8. Sankey Diagram (Highlighting Only Largest Gains)

10.8.1 Analysis: Sankey Diagram: Only Largest Gains

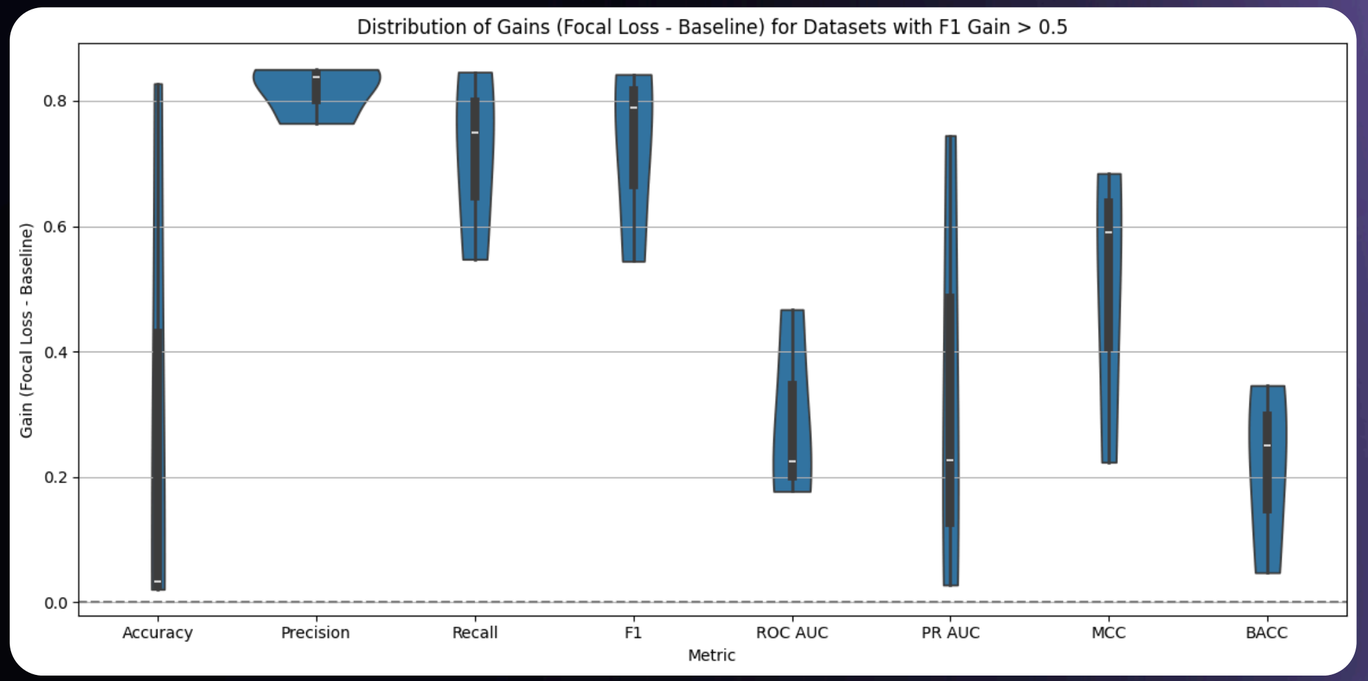
- **Widespread Gains:** Many datasets show substantial improvements in at least one metric, and several datasets have gains in multiple metrics.
- **Most Common Gains:** The thickest and most frequent links are to **Precision**, **F1**, and **Recall**. This means Focal Loss most often delivers large improvements in these metrics, which are critical for imbalanced classification.
- **Less Frequent Gains:** Fewer datasets show large gains in **Accuracy**, **MCC**, **BACC**, **ROC AUC**, and **PR AUC**. This is expected, as these metrics are harder to improve substantially, especially in highly imbalanced settings.
- **Dataset Diversity:** The improvement is not limited to a single dataset type; gains are distributed across many different datasets, indicating Focal Loss can be broadly beneficial.

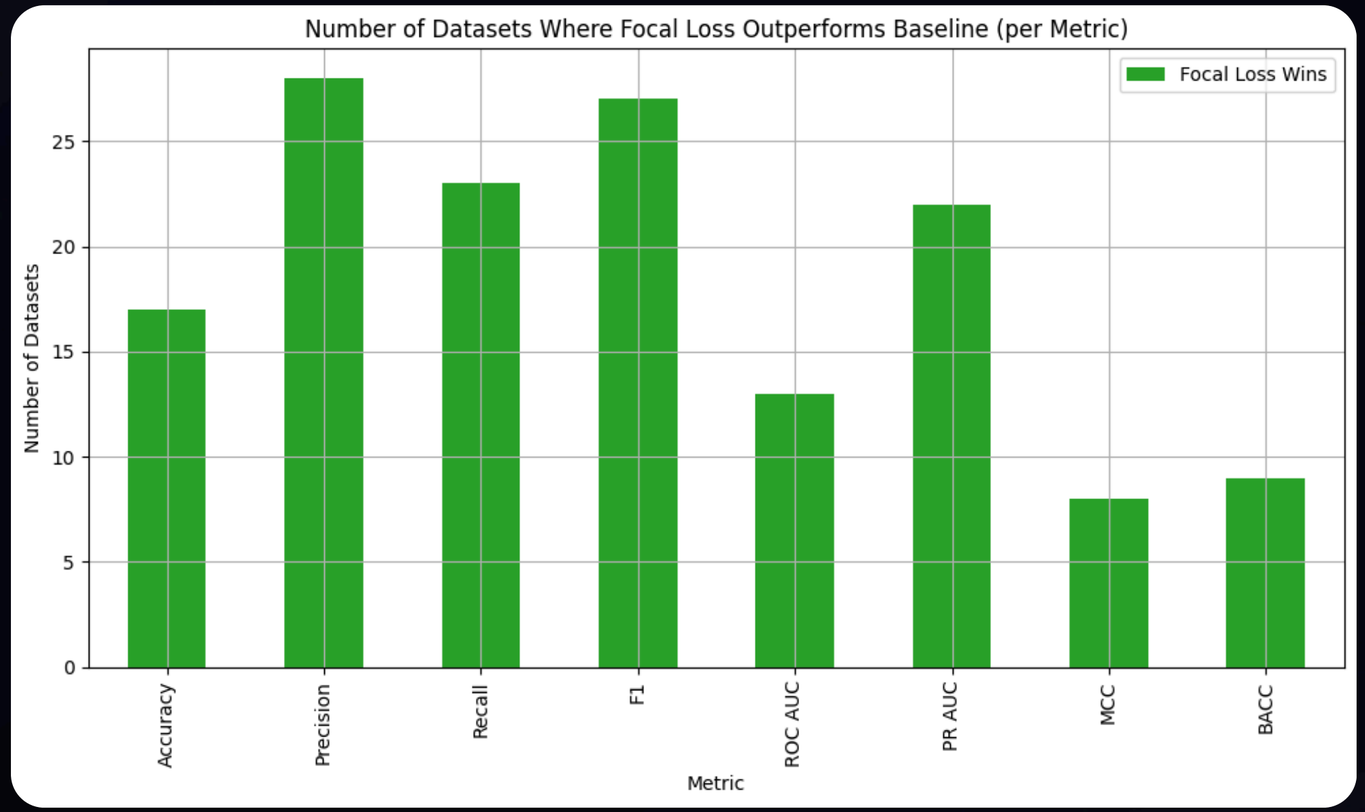


10.9. Violin Plot of Differences (Highlighting Specific Datasets)

10.9.1 Anlysis: Distribution of Gains (Focal Loss - Baseline) for Datasets with F1 Gain > 0.5

For datasets where Focal Loss delivers a large F1 improvement, it also tends to substantially boost Precision and Recall, and to a lesser extent, other metrics. This highlights Focal Loss's strength in improving minority class detection and overall balanced performance in the most challenging cases. However, the impact on metrics like ROC AUC and BACC, while positive, is less dramatic and more dataset-dependent.





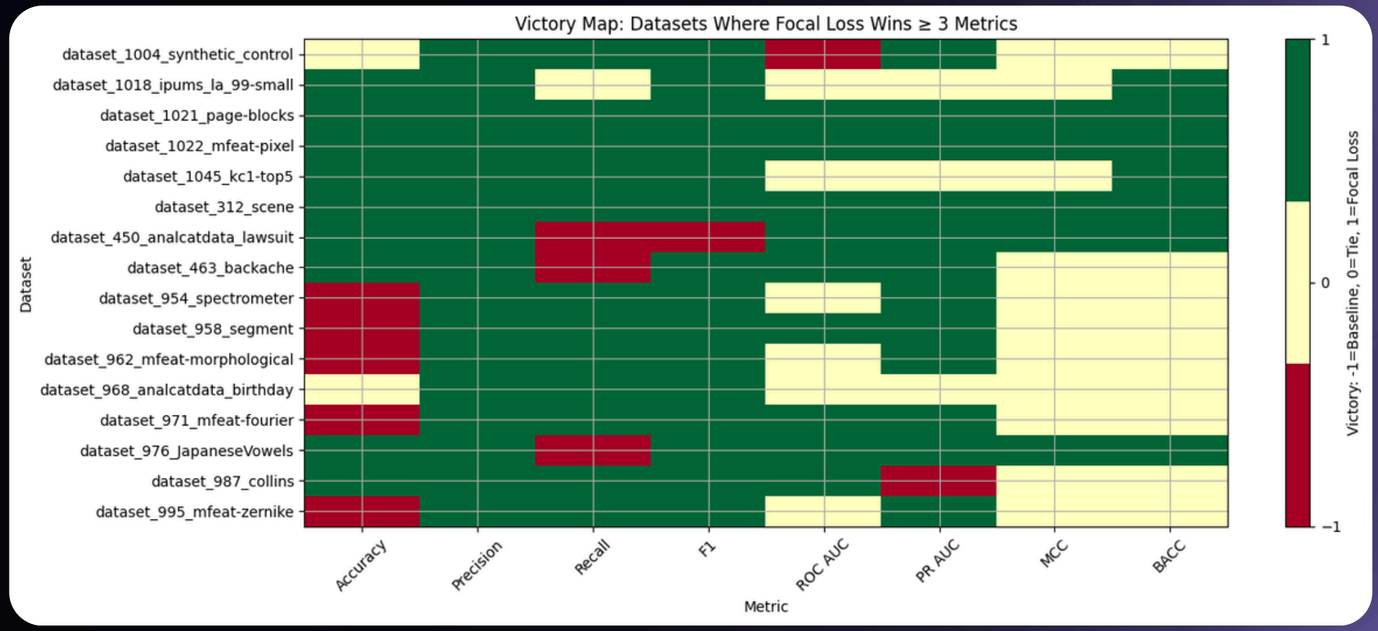
10.10 Stacked Bar Chart of "Wins" per Metric (Highlighting Only Wins)

10.10.1 Analysis: Stacked Bar Chart of "Wins" per Metric (Highlighting Only Wins)

- Focal Loss outperforms Baseline most often in Precision, F1, and Recall (over 20 datasets each).
- F1 and Precision show the highest number of "wins," confirming Focal Loss is especially effective for minority class detection and balanced performance.
- ROC AUC, PR AUC, Accuracy: Moderate improvement, but less frequent.
- MCC and BACC: Focal Loss rarely wins, indicating limited gains in overall balanced metrics.

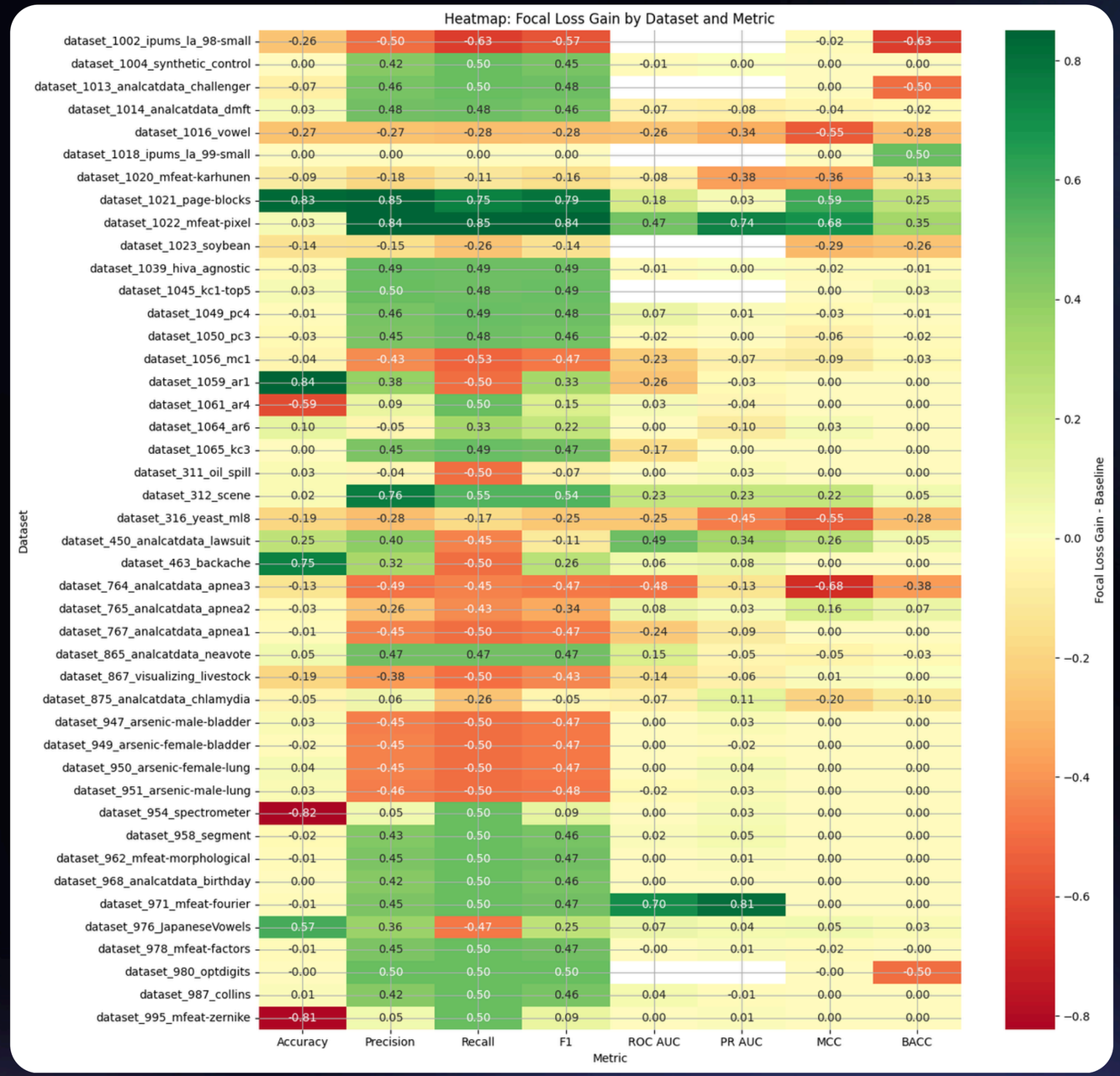
10.11. "Victory Map" per Dataset and Metric (Highlighting Specific Datasets)

- Green cells (1): Focal Loss wins that metric for the dataset.
- Yellow (0): Tie; Red (-1): Baseline wins.
- Most datasets show Focal Loss winning in ≥ 3 metrics → rows mostly green.
- Strong advantage in Precision, Recall, and F1 → almost entirely green columns.
- A few exceptions (e.g., datasets 954, 971) where Baseline wins some metrics.
- Ties appear mainly in ROC AUC, PR AUC, MCC, and BACC → harder to consistently improve.
- Conclusion: Focal Loss often wins broadly across metrics, confirming its effectiveness for imbalanced datasets, despite a few isolated exceptions.



10.12 Heatmap: Focal Loss Gain by Dataset and Metric

- Each cell shows Focal Loss - Baseline; green = gain, red = decline.
- Precision, Recall, F1: Show the most consistent gains → better minority class detection.
- Recall and F1 improve especially on heavily imbalanced datasets.
- ROC AUC, PR AUC: Mixed results → gains in some datasets, losses in others.
- Accuracy: Moderate, dataset-dependent gains, due to majority class influence.
- MCC and BACC: Limited improvement, suggesting better detection doesn't always mean better balance.
- Some datasets show strong multi-metric gains (e.g., 1021, 1022), while others show declines (e.g., 954, 995).
- Conclusion: Focal Loss generally improves key metrics for the minority class, but its effectiveness varies by dataset and metric.



Practical Assignment ML1

Logistic Regression

...

on Imbalanced Data

Baseline Model

- High accuracy, but biased toward the majority class
- Very poor minority class metrics (Precision, Recall, F1)
- High variability across datasets

Focal Loss Model

- Improves Recall, Precision, and F1 for the minority class
- More consistent across datasets
- Especially effective in severely imbalanced cases
- Mixed results on ROC AUC, PR AUC, MCC, BACC

Key Insights

- Better class balance (confusion matrix)
- Statistically significant gains in Precision and F1
- Effectiveness depends on dataset; requires gamma tuning

Final Takeaway and Future Work

Focal Loss is a valuable improvement over the baseline, but not a one-size-fits-all solution, tuning and careful evaluation remain essential.Future work includes testing more advanced imbalance techniques (e.g., resampling, ensembles) and automating hyperparameter tuning for Focal Loss. Applying the approach to real-world datasets and expanding model interpretability are also recommended.