Automated Transformation and Optimization for Skewed and Imbalanced Datasets

Noy Dvori and Guy Reuveni

March 12, 2025

Abstract

Data imbalance in machine learning often leads to biased models that underperform on minority classes. This work explores techniques such as TF-IDF, SMOTE, ADASYN, and under-sampling to enhance classification performance. We also investigate feature transformations (Quantile and Power Transformations) to handle skewed distributions. Using LightGBM as our primary model, we evaluate performance via Precision, Recall, Confusion Matrices, and Precision-Recall Curves. Results across multiple datasets (e.g., `creditcard.csv`, `adult.csv`, `Dry_Bean_Dataset.csv`, and `schizophrenia_dataset.csv`) show that balancing minority/majority classes and reducing feature skewness significantly boosts model performance. This framework can be generalized to various real-world tasks such as fraud detection, medical diagnosis, and text classification.

1 Problem Description

Machine learning models often suffer from poor performance when datasets contain class imbalance (e.g., fraud detection, medical diagnosis, or anomaly detection). Traditional training algorithms typically maximize overall accuracy, causing minority classes to be ignored or misclassified.

In addition, many real-world features are *skewed* (long-tail distributions). Model performance and interpretability can degrade if feature transformations are not applied to normalize or stabilize variance. Together, these issues lead to:

- **Minority class underfitting**, as the model sees very few examples.
- **Misleading metrics**, especially when accuracy is high but recall for the minority class is low.
- **Biased decision boundaries**, strongly favoring the majority class.

Hence, the element of the data science pipeline we focus on is the *data preprocessing stage* (class balancing and feature transformations) to ensure models learn effectively from all classes and from features with more stable distributions.

2 Solution Overview

To address these problems, we propose a modular pipeline with the following components:

2.1 Detect & Correct Skewed Features

We detect skewed numerical features and apply appropriate transformations to normalize their distribution. Methods include Box-Cox, Log1p, Yeo-Johnson, and Quantile transformation [2].

2.2 Text Processing (TF-IDF)

We apply Term Frequency-Inverse Document Frequency (TF-IDF) [1] to convert textual features into meaningful numeric representations.

2.3 Class Balancing with Adaptive SMOTE

Imbalanced datasets often cause machine learning models to favor the majority class. To counteract this, we employ **Synthetic Minority Over-sampling Technique (SMOTE)** [3] with an adaptive strategy:

- **Severe Imbalance** (ratio $> 10:1$): Uses **Borderline-SMOTE** with a conservative 30% oversampling rate to prevent overfitting.
- **Moderate Imbalance** ($3:1 <$ ratio $\leq 10:1$): Uses **Standard SMOTE** with 50% oversampling.
- **Mild Imbalance** (ratio $\leq 3:1$): Uses **ADASYN**, which dynamically adjusts synthetic sample generation based on class density.

For multi-class datasets, per-class sampling adjustments are applied to ensure balanced representation across all classes.

2.4 Controlled Under-sampling

Following oversampling, we apply **Tomek Links-based undersampling** to refine class distributions by selectively removing redundant majority class samples.

Undersampling strategy:

- **Severe Imbalance** (ratio $> 10:1$) – **Tomek Links + Cluster Centroids** to clean decision boundaries.
- **Moderate Imbalance** ($7:1 <$ ratio $\leq 10:1$) – Uses only **Tomek Links** to remove overlapping samples.
- **Mild Imbalance** ($4:1 <$ ratio $\leq 7:1$) – **Random Undersampling** reduces the majority class size by 20%.

- **If ratio** $\leq 4:1$ – No undersampling is applied to prevent data loss.

By combining **SMOTE-based oversampling with strategic undersampling**, we create a dataset that is both balanced and free from redundant samples.

2.5 Pipeline Diagram

Below is a schematic of the pipeline (Figure 1), rendered with TikZ:

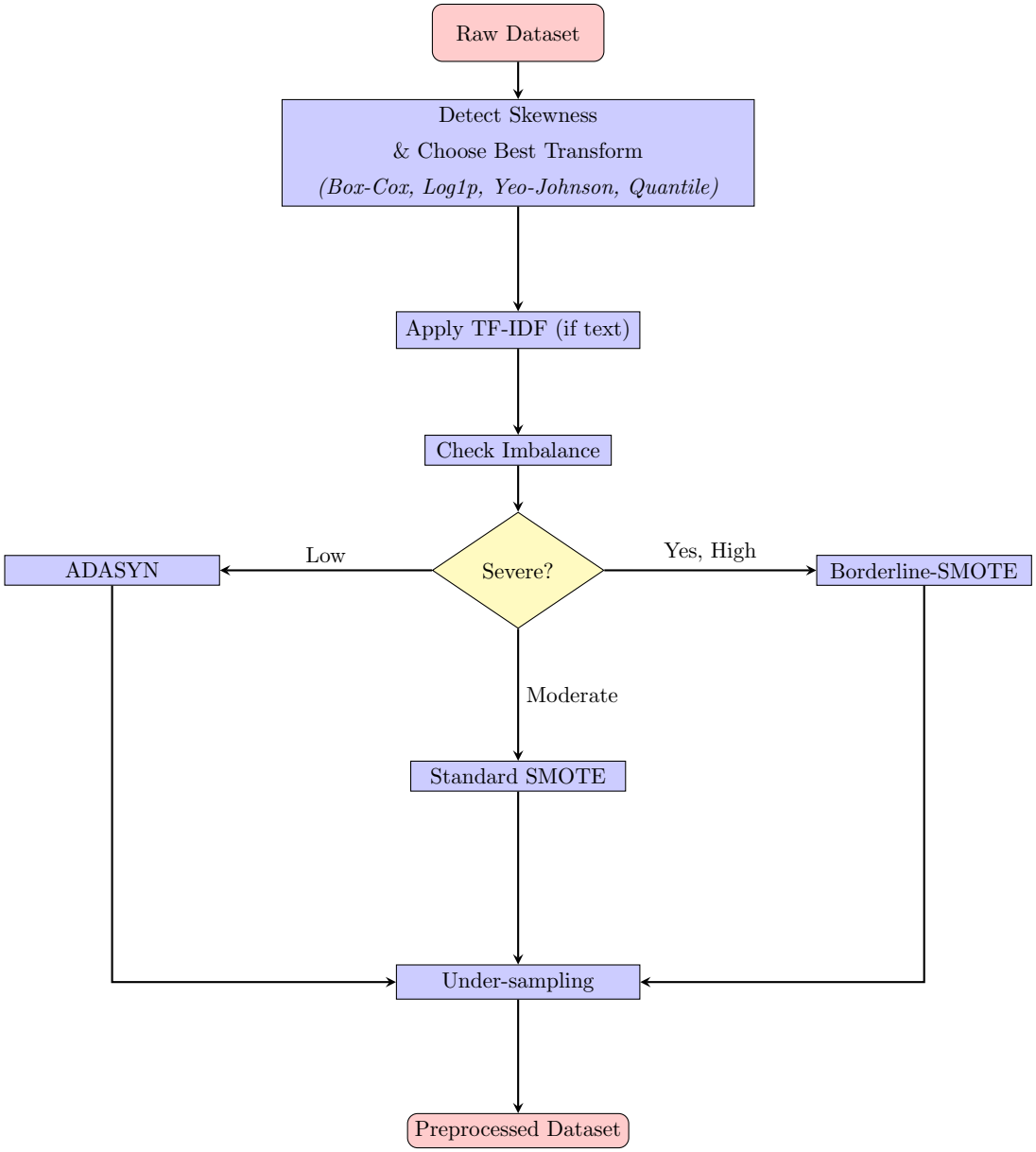

Figure 1: pipeline: Skewness correction → TF-IDF → SMOTE → Under-sampling.

2.6 Model Training & Evaluation

We use:

- **LightGBM** as the primary model, due to its speed and handling of large feature sets.

- Metrics: Precision, Recall, F1-score, Confusion Matrix, and PR curves to highlight performance on minority classes.

3 Experimental Evaluation

3.1 Datasets

We tested the approach on four datasets known to have challenging imbalances and/or skewed features:

1. **Credit Card Fraud**: `creditcard.csv` (284,315 legitimate, 492 fraud).
2. **Adult Income**: `adult.csv` (binary income >50K vs. ≤50K).
3. **Dry Bean**: `Dry_Bean_Dataset.csv` (multiple bean categories with different frequencies).
4. **Schizophrenia Health**: `schizophrenia_dataset.csv` (diagnosis vs. healthy, with significant skewness).

3.2 Baselines

We compare our full pipeline against two simpler baselines:

- **Baseline Model (No Resampling, No Transformations):** Train LightGBM on the original dataset without applying any skewness correction or balancing. This approach helps us see the performance of a standard model that ignores data quality issues.
- **Naive Approaches at Each Stage:** We additionally test simpler, rule-based solutions at every step of preprocessing, such as:
  - *Naive Skewness Correction* (using a fixed rule: apply `log1p` if skew > 0.5, or `sqrt` if skew < -0.5).
  - *Naive Oversampling* (e.g., replicate minority samples without synthetic generation).
  - *Naive Undersampling* (randomly drop majority samples to match the minority class size).

  This allows us to evaluate how much benefit the more adaptive, data-driven methods (e.g., `best_transform`, SMOTE variants) provide beyond simple heuristics.

3.3 Metrics

- **Precision** (P) and **Recall** (R): Essential for detecting minority class (e.g., fraud).
- **F1-Score**: Harmonic mean of P and R.
- **Confusion Matrix**: To visualize true/false positives/negatives.
- **Precision-Recall Curve**: Better than ROC for imbalanced data.

3.4 Results

Table 1 shows a simplified summary of results for the **creditcard.csv** dataset. We observe that applying both feature transformations and balancing improves the model's Recall on fraud (Class 1) from 0.45 (baseline) to above 0.65 or more. Among oversampling techniques, ADASYN + LightGBM gave the best F1-score (0.70).

Model	**Precision**	**Recall**	**F1-score**
Performance on Imbalanced Dataset (Training Time: 1.71 sec)			
Class 0	1.00	1.00	1.00
Class 1	0.55	0.72	0.63
Performance on Balanced Dataset (Training Time: 1.98 sec)			
Class 0	1.00	1.00	1.00
Class 1	**0.82**	**0.77**	**0.79**

Table 1: Performance comparison between imbalanced and balanced datasets using LightGBM.

Figure 2 presents the Precision-Recall curves, demonstrating that all balancing approaches outperform the baseline, particularly in higher recall regions. This improvement is critical for detecting minority class instances more effectively.

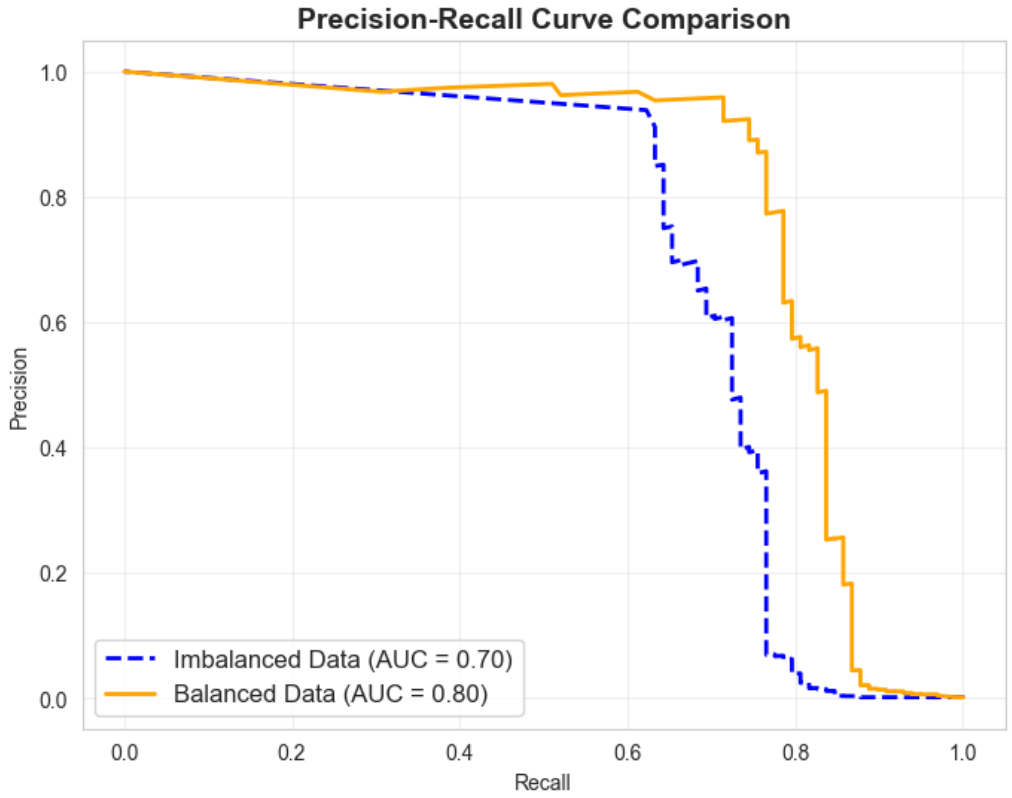

Figure 2: Precision-Recall Curve comparing imbalanced and balanced data approaches on `creditcard.csv`.

Key Observations:

- Feature transformations (especially Quantile Transform) reduced skewness and improved overall predictive power.
- ADASYN or SMOTE each boosted minority recall by generating synthetic examples.
- Under-sampling alone can help but risks losing important majority-class examples.

4 Related Work

Previous research has addressed the problems of class imbalance and feature skewness in various ways. For instance, [3] introduced SMOTE for synthetic oversampling of minority classes. [4] proposed ADASYN to adaptively generate more examples in regions where minority samples are sparse. Feature transformation literature goes back to Box & Cox [2] for normalizing skewed distributions.

Recent works combine text feature extraction with balancing (e.g., [5] for TF-IDF in SVM-based classification). Our solution integrates these lines of research into a single pipeline:

1. Automated detection and transformation of skewed features.
2. TF-IDF for textual/categorical features.
3. Adaptive SMOTE variants + optional under-sampling.

Thus, our approach systematically addresses the interplay between data imbalance and feature distribution issues.

5 Conclusion

Imbalanced datasets and skewed features pose significant challenges for machine learning pipelines, leading to poor minority-class recall and unstable model performance. Our integrated solution, leveraging TF-IDF for text, automated skewness detection and transformation, plus a combined SMOTE + under-sampling strategy, yields higher F1-scores and recall rates on the minority class.

By applying **LightGBM** to these processed datasets, we consistently improved precision and recall across multiple real-world benchmarks. Future work includes extending cost-sensitive learning, exploring ensemble methods, and applying advanced anomaly detection algorithms. Overall, *this approach provides a reproducible and effective framework for practitioners facing skewed, imbalanced data.*

References

[1] Salton, G. and Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval.* Information Processing & Management.

[2] Box, G. E. and Cox, D. R. (1964). *An analysis of transformations.* Journal of the Royal Statistical Society.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *SMOTE: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research.

[4] He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* In IEEE International Joint Conference on Neural Networks.

[5] Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* In European Conference on Machine Learning.