Intervening in problematic data regions to improve machine learning models

Gregully Willian[1], Fabio Porto[2], Eduardo H. M. Pena[1]

[1] Federal University of Technology - Paraná,
Campo Mourão - PR - Brazil

[2]National Laboratory of Scientific Computing (LNCC)
Petrópolis - RJ - Brazil

gregullwilliany@gmail.com, fporto@lncc.br, eduardopena@utfpr.edu.br

Abstract. *Debugging machine learning models is essential to improving their robustness and performance. This work explores a data-driven debugging approach based on problematic data regions—subsets of the data where the model performs poorly compared to others. These regions often reflect problems such as class imbalance, bias, or unfairness. We propose to improve model performance by focusing primarily on the data, using a specialized algorithm to identify problematic regions in the datasets, and then applying targeted interventions. A common cause for poor performance is class imbalance within a problematic region. For such scenarios, we apply data augmentation in a controlled manner, avoiding excessive introduction of synthetic data. Our experiments demonstrate that it is possible to improve model performance by focusing exclusively on problematic data regions rather than the entire dataset.*

1. Introduction

Machine learning (ML) models are increasingly being deployed across various fields—from agriculture to healthcare—where their decisions can carry profound consequences [Shehab et al. 2022, Sharma et al. 2021]. One practical way to work toward making models robust and reliable is through model debugging—looking closely at how a model responds to different inputs and identifying where and why it fails. When we understand the mistakes a model tends to make, we can take more targeted steps to improve its performance. Principled practices for model debugging include hyperparameter tuning, error analysis, and checking for overfitting or underfitting [Yang and Shami 2020]. While these methods are helpful, they often overlook deeper issues arising from *how the model interacts with the data*. To address these limitations, researchers have developed tools for better understanding the relationship between models and data—such as *structured explanation formats* [El Gebaly et al. 2014] and, more recently, algorithms for detecting *problematic data regions* [Chung et al. 2020, Sagadeeva and Boehm 2021, Kerrigan and Bertini 2023].

Some of the limitations in standard model evaluation motivate the identification of problematic data regions. Model performance is typically assessed using global metrics such as precision, recall, F1 score, or log loss [Sagadeeva and Boehm 2021]. While these metrics offer a high-level view of effectiveness, they often obscure issues affecting specific subgroups in the data, such as bias, unfairness, or imbalance. In this context, detecting problematic regions becomes essential. A problematic data region—a data slice—is a subset of the dataset defined by a conjunction of feature values (e.g., *age group* = 18-25 $\wedge$ *account type* = basic). A slice is problematic if the model's performance, measured by a loss function like log loss, differs significantly from that of a reference group, often the slice's complement [Chung et al. 2020]. This discrepancy reveals areas where errors are more concentrated, an issue that is not easily captured by aggregate metrics such as F1 score [Pastor et al. 2021, Kerrigan and Bertini 2023]. This work focuses on such problematic regions as a way to guide targeted interventions—particularly data-related adjustments—to improve model performance.

As an example of problematic data regions automatically identified, consider the Census Income dataset [Kohavi 1996] and the regions shown in Table 1, discovered using a specialized algorithm [Chung et al. 2020] (we describe such tools later). A random forest model trained for this dataset exhibits substantially worse performance in these regions, as reflected by elevated log loss values compared to the overall dataset. Log loss measures the accuracy of predicted probabilities in classification tasks, with higher values indicating greater uncertainty or incorrect predictions. The extent of this degradation is quantified using *effect size*, which normalizes the difference in loss between the data region and its complement and helps to make comparisons across regions of varying sizes.

Table 1. Data regions automatically identified in the Census Income dataset, with relatively high log loss and effect size.

Region	Log Loss	Size	Effect Size
All	0.30	32,561	n/a
Marital Status = Married	0.55	14,065	0.60
Relationship = Husband	0.55	12,463	0.55
Country = US $\wedge$ *Relationship* = Wife	0.60	1,251	0.40
Capital Gain = 0	0.64	1,230	0.46

Previous work has extensively investigated identifying problematic regions and their statistical divergence from the overall dataset [Chung et al. 2020, Sagadeeva and Boehm 2021, Zhang et al. 2023, Pastor et al. 2021]. However, there remains a gap in translating these insights into concrete strategies for improving model performance. Existing methods typically leave it to the practitioner to determine how best to respond to these findings, offering little guidance on next steps.

In this work, we address this gap by proposing a methodology that treats problematic data regions as intervention targets—guiding data-centric actions such as data augmentation to mitigate performance issues and enhance model quality. Rather than applying data-centric actions indiscriminately across the dataset, we concentrate on problematic regions that are automatically identified, where we know the model exhibits significant performance degradation. The goal is to enhance model accuracy while minimizing the risk of overfitting or introducing noise in well-performing data regions.

We use a tool called `Slice Finder` to discover interpretable problematic regions due to its practical utility, relatively low computational cost, and ease of integration into data processing workflows [Chung et al. 2020]. We observed that class imbalance is a recurring source of poor predictive performance within these regions. To mitigate this, we employ a localized oversampling approach inspired by SMOTE [Chawla et al. 2002], synthesizing new examples by interpolating between minority class samples. We evaluate the effectiveness of localized oversampling by comparing it to the global application of SMOTE and the baseline model, measuring its impact on important performance metrics such as precision, recall, and F1 Score. The contributions of this work are as follows:

1. We propose region-aware data balancing strategies that improve model accuracy;
2. We analyze the impact of targeted interventions in problematic data regions on end-to-end model performance;
3. We present case studies demonstrating how automated detection of problematic regions enables effective data debugging and targeted intervention;
4. We evaluate our approach across real-world datasets from varied domains, using a state-of-the-art tool for problematic data region detection.

2. Background

We follow the notation from [Chung et al. 2020]. Let D be a dataset with n instances, each instance $x^{(i)}$ described by a set of features $\mathcal{F} = \{F_1, F_2, \ldots, F_k\}$. Features may be categorical (e.g., *education* with values like Bachelor's, Master's) or discretized numerical intervals (e.g., *age* in $[0, 5]$). Each instance is associated with a label $y^{(i)}$, giving us a dataset $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$. Let h be a trained model that maps feature vectors x to predicted labels. Our goal is to assess how h performs on different subsets of the data, which we refer to as *regions*. We use a conjunction of feature-based conditions to define a data region: $\bigwedge_j F_j$ op v_j, where F_j are distinct features, $op \in \{=, \leq, \geq, >, <, \neq\}$, and v_j are constants. For example, a region *age* $> 18 \wedge$ *education* = Master's.

We use two key metrics to evaluate the model's performance in a data regions: *logarithmic loss* (log-loss) and *effect size* [Chung et al. 2020]. Log-loss measures the distance between predicted probabilities and true labels in binary classification tasks, ranging from 0 (perfect prediction) to 1 (poor prediction), with lower values indicating better performance. The effect size complements this by quantifying the magnitude of the performance difference between a region and the rest of the dataset. Unlike statistical significance, which large sample sizes can skew, effect size reflects whether this difference is meaningful [Cohen 1988]. It is calculated by standardizing the loss difference between the region and its complement using their standard deviations. This helps distinguish between negligible and substantial disparities [Chung et al. 2020]. Standard thresholds define values around 0.2 as small, 0.5 as medium, 0.8 as large, and 1.3 as very large [Cohen 1988, Chung et al. 2020]. In our context, regions with larger effect sizes indicate areas where the model consistently underperforms and should be prioritized for intervention, even if the region is relatively small.

3. Intervention strategies for problematic data regions

3.1. Overview of region-aware interventions

We observed that class imbalance within specific data regions plays a major role in degrading model performance. In our tested datasets, several problematic regions were commonly dominated by a single class to a great degree, resulting in biased predictions. To address this, we propose a targeted data augmentation strategy to strengthen the minority class representation in these regions. We also explored a more aggressive intervention by removing entire problematic regions to isolate their impact on model accuracy. Building on these observations, we leverage the identified region to detect the features most strongly associated with generalization issues.

3.2. Problematic data region discovery

Following the setup from [Chung et al. 2020], the problematic data region detection problem is to automatically discover data regions where model h performs significantly worse than elsewhere in the dataset. For any region $s \subseteq D$, we have its complement as $s' = D \backslash s$. The relative performance drop is measured by the difference in log-loss ψ between s and S': $\psi(s, h) - \psi(s', h)$. If this value is positive and statistically significant, s is flagged as a problematic region. Each region is treated as a hypothesis and must satisfy two main criteria: the loss in the region s must be statistically significantly higher than in its complement s', and the magnitude of this difference—measured by effect size—must be large enough to be practically meaningful. Given a desired number of regions k, a minimum effect size threshold θ, and a significance level α, the goal is to find the top k regions that meet the following conditions: (i) the effect size is greater than or equal to θ; (ii) the statistical significance level is less than α; and (iii) no strictly smaller subset of predicates defines a better-performing region under the same criteria. To enhance interpretability and data coverage, the data regions are prioritized based on (i) fewer predicates, (ii) a larger instance count, and (iii) a greater effect size.

To detect such regions, we use `Slice Finder`, a tool that implements two complementary methods: decision tree partitioning and lattice-based search [Chung et al. 2020]. The decision tree method builds a classifier trained to separate misclassified from correctly classified examples. It constructs binary splits that minimize impurity, producing a tree whose leaf nodes correspond to candidate regions. These leaf regions are evaluated breadth-first for effect size and statistical significance. If a leaf exceeds a predefined effect size threshold θ, the algorithm attempts to generalize by moving upward in the tree. While decision trees offer interpretable structure, they are limited in that splits are optimized for prediction rather than error analysis, and tree partitions are mutually exclusive, preventing detection of overlapping problematic regions.

The lattice-based approach addresses these limitations by exploring the full combinatorial space of feature-value conditions. It organizes slices into a lattice, where each node represents a conjunction of feature predicates. The method performs a breadth-first search, starting with single-feature slices and gradually refining them by adding more conditions. Numerical features are discretized in advance, and for categorical features with high cardinality, only the top-N most frequent values are considered. Because such exhaustive search can produce false positives due to multiple comparisons [Foster and Stine 2008], `Slice Finder` applies marginal false discovery rate to ensure that only statistically sound regions are reported, even in large search spaces.

3.3. Proposed approach

While prior work focuses on discovering problematic regions, our goal in this work is to use these regions to improve model performance through targeted data interventions. Given a model h trained on dataset D, we aim to produce an improved model h' trained on a modified dataset D'. First, we (automatically) detect a set of regions S in which h

performs poorly. We then apply targeted interventions—such as data balancing or augmentation—within these regions to construct D'. Finally, we train a new model h' on this updated dataset and evaluate whether its performance improves. This approach allows us to address performance bottlenecks precisely where they matter most without modifying the entire dataset or retraining the model mindlessly. Instead of treating poor performance as a black-box artifact, we treat it as an actionable signal for focused model improvement.

3.3.1. Local Balancing with SMOTE

To evaluate the effectiveness of targeted data balancing within problematic regions, we conducted experiments comparing two data augmentation strategies: (i) *slice-specific augmentation*, which applies SMOTE exclusively within problematic regions to balance local class distributions, and (ii) *full-dataset augmentation*, which applies SMOTE globally across the entire dataset. We compared the performance of models trained on the augmented datasets against their baseline counterparts, using standard evaluation metrics such as precision, recall, and F1-score.

We formalize the local balancing procedure as follows. Let $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ be the dataset, where each $x^{(i)}$ is described by a set of features $\mathcal{F} = \{F_1, F_2, \ldots, F_k\}$ and associated with a label $y^{(i)}$. Let $S = \{s_1, s_2, \ldots, s_m\}$ be the set of m problematic regions identified by `Slice Finder`, where each $s_i \subseteq D$ is defined by a conjunction of feature-value conditions over $\mathcal{F}$.

For each region $s_i \in S$, we compute its class imbalance ratio

$$r_i = \max\left(\frac{|s_i^+|}{|s_i^-|}, \frac{|s_i^-|}{|s_i^+|}\right),$$

where s_i^+ and s_i^- denote the subsets of positive and negative instances in s_i, respectively. A region is selected for augmentation if its imbalance ratio r_i exceeds a threshold τ.

Once selected, SMOTE is applied within each region by interpolating between minority class instances and their nearest neighbors to generate synthetic samples. The augmented region s_i' replaces s_i in the dataset, resulting in a modified dataset D'. Such targeted interventions enable localized performance improvements without significantly disturbing the overall data distribution, allowing the model to better generalize in regions previously prone to biased predictions.

3.3.2. Evaluating region impact via slice removal

Problematic regions can significantly degrade model performance and, if not adequately addressed, may invalidate its predictions [Chung et al. 2020]. While removing these regions might seem like an attractive preprocessing strategy, it risks discarding valuable information, compromising the model's generalization ability, and leading to overfitting. Although metric improvements may be observed after removal, they often do not translate to better real-world performance. Nevertheless, analyzing the effect of removing problematic regions offers valuable insights. It highlights their contribution to model behavior and helps identify feature patterns associated with poor performance. To this end, we evaluate the impact of excluding problematic regions compared to randomly removing regions of equivalent size.

3.3.3. Intervention pipeline with feature-based analysis of problematic patterns

While preliminary findings suggest that internal class imbalance is a major contributor to poor performance, other distributional factors may also negatively affect model behavior. For example, regions with skewed or sparse feature distributions can hinder consistent pattern learning. We systematically conduct case studies that combine statistical analysis with targeted interventions to explore these characteristics. We designed an intervention pipeline based on the insights from feature-based analysis [Liu and Motoda 1998].

We began with a *predicate frequency analysis*, counting the occurrence of each feature within the problematic regions. We extracted the features from the region data, computed their frequencies, and then ranked them in ascending frequency. Next, we conducted a *univariate distribution analysis* focusing on features that appeared in up to 30% of the problematic regions. We examined the mean, number of samples, and standard deviation for these features and used histograms to visualize their distributions better. We experimented with different thresholds for the percentage of regions considered problematic and found that the used threshold could identify significant patterns and maintain focused analysis. Lower thresholds tended to produce more scattered and less meaningful regions, making it harder to draw consistent conclusions.

In the stage of *pattern identification*, we observed recurring characteristics among problematic regions, particularly continuous attributes with highly dispersed values and sparse distributions across a broad range. We then applied *targeted binning* to continuous features with high cardinality. Discretization was applied if more than 10% of the values were unique relative to the total number of samples. We tested several threshold values in preliminary experiments and found that the chosen number consistently yielded better model performance and interpretability results. Then, we performed a *performance evaluation* using five-fold cross-validation to assess the effect of each intervention, tracking changes in the F1-score as the primary performance metric. Finally, we conducted *residual detection and data augmentation* by reapplying the problematic region detection process to uncover the remaining issues. For these, we applied our general balancing strategy to address any residual class imbalances.

4. Experimental evaluation

4.1. Experimental Setup

Dataset. Table 2 summarizes the main characteristics of the datasets used in our experiments. These datasets involve different domains and scenarios, varying in size and degrees of class imbalance. All involve binary classification tasks with both categorical and numerical attributes. The Census Income dataset [Kohavi 1996] contains demographic and income information to predict whether an individual's annual income exceeds \$50,000. The Bank Marketing dataset [Moro et al. 2012] includes data from direct marketing campaigns via telephone, aiming to predict whether a customer will subscribe to a term deposit. Lastly, the Online Shop dataset contains customer browsing data to predict purchase intentions in an e-commerce setting [Sakar and Kastro 2018].

Table 2. Datasets used for experiments

Dataset	Instances	Features	Positive Class	Negative Class
Census Income (Census)	30,162	14	7,508	22,654
Bank Marketing (Bank)	45,211	16	5,289	39,922
Online Shop (Shop)	12,330	18	1,908	10,422

We use abbreviations to represent problematic data slices identified in our experiments for clarity. The most frequent slices with their corresponding attributes, values, and encoder information are listed in Table 3.

Table 3. Problematic data slices identified in the experiments with their corresponding attributes, values, and encoder indices.

Abbreviation	Attribute	Value	Encoder
MS	*MaritalStatus*	Married-AF-spouse; Never-married; Separated	2, 4, 5
RS	*Relationship*	Wife	5
EDU	*Education*	Prof-school	14
SEX	*Sex*	Male	1
CG	*Capital Gain*	(numerical)	–
ADM	*Administrative*	(numerical)	–
PR	*ProductRelated*	(numerical)	–

Models and training. We split the dataset into 80% for training and 20% for validation (hold-out). We use the Random Forest model with hyperparameter optimization through BayesSearchCV. The model is trained with stratified cross-validation (k=5) on the training set, with shuffling for greater robustness. In the proposed approach, we identify the problematic regions exclusively in the training set, perform targeted balancing on these regions, and train a new model. The new evaluation is done on the validation set, again applying cross-validation (k=5) on the training to estimate the performance before testing on the hold-out.

Baselines and Implementation. For the baselines, we applied SMOTE oversampling [Chawla et al. 2002] to the entire dataset, while our approach focuses only on problematic regions. We adopted as an intervention criterion a degree of imbalance greater than 30% between classes within the region, a value that has proven to be adequate in the experiments performed. We temporarily isolated the region, subjected it to the SMOTE oversampling process, and reintegrated it into the training set. After that, the model was retrained and evaluated on the validation set.

SMOTE was configured to fully balance the minority class, using its default variant with predefined parameters. By default, synthetic sample generation is based on the five nearest neighbors of each minority class sample. Although we briefly explored variations of the algorithm, we chose to maintain the default configuration to ensure the generalizability of the results. No specific SMOTE variant was adopted as a basis for the experiments since the problematic regions presented distinct characteristics, making it unfeasible to define a single configuration that would suit all of them. Finally, we highlight that this global balancing approach strongly impacted correcting the imbalance. However, it also increased the risk of introducing synthetic noise — an effect we seek to mitigate with our localized balancing proposal.

`Slice Finder` was used as proposed by the original authors. A Random Forest is trained to estimate classification errors and identify underperforming regions based on statistical significance tests and effect size. Regions are generated by binary splits into numeric, categorical, or binary attributes using a decision tree with limited depth (default: 2), which seeks to preserve interpretability and reduce overfitting when segmenting the attribute space. The number of regions returned is adjustable, but we followed the examples provided by the authors, limiting the worst-performing regions to 100. These regions are ordered based on effect size, log-loss, number of samples, or another metric of interest. The minimum effect size for a region to be considered problematic was set at 0.4, as suggested by the authors. This value is based on Cohen's classification, in which values above 0.4 indicate a moderate to strong effect—allowing us to focus on regions with substantial deviations in performance [Cohen 1988].

Hardware. All our experiments were performed on a machine AMD ryzen 7 mobile 3700U with 12 GB memory RAM and 4 CPU cores with 2.3 GHz clock speed.

4.2. Results with target interventions data augmentation

Based on the hypothesis that class imbalance is a key factor for poor performance in problematic regions, we evaluate the models in three settings: (i) slice-specific augmentation, where SMOTE is applied exclusively to the problematic regions to balance class distributions; (ii) full-dataset augmentation, where SMOTE is applied to the entire dataset; and (iii) no SMOTE, where no augmentation is applied. The results are presented in Table 4.

Table 4. Model performance comparison on hold-out validation set: baseline (no augmentation), conventional SMOTE (full-dataset), and our approach (targeted SMOTE in problematic regions).

Set/Region	Instances	Accuracy	Recall	F1 Score
Bank				
Bank (No SMOTE)	36,168	0.79	0.69	0.73
Bank + SMOTE	63,874	0.72	0.77	0.75
poutcome = sucess + SMOTE	36,521	0.79	0.71	0.74
month = October + SMOTE	36,242	0.79	0.73	0.76
Census				
Census (No SMOTE)	30,162	0.88	0.83	0.85
Census + SMOTE	45,308	0.91	0.93	0.92
MS = 2 + SMOTE	31,429	0.86	0.82	0.84
RS = Husband + SMOTE	31,267	0.90	0.87	0.88
CG = 0 $\wedge$ RS = Husband + SMOTE	30,264	0.92	0.89	0.90
Shop				
Shop (No SMOTE)	9864	0.84	0.76	0.79
Shop + SMOTE	16676	0.78	0.81	0.79
adiministrative = 13 + SMOTE	9881	0.83	0.78	0.80
PR = 58 + SMOTE	9882	0.83	0.77	0.80

The results reveal that targeted balancing of class distributions in problematic regions often achieves relevant performance improvements while maintaining the overall structure of the dataset in non-problematic areas. The approach was effective in both the Shop and Bank datasets, where targeted SMOTE in specific regions produced higher F1 scores relative to full augmentation while requiring minimal data expansion. This approach demonstrates particular promise in the Census dataset, where augmentation of the "CG = 0 $\wedge$ RS = Husband" region achieved good performance (F1 score of 0.90) with substantially fewer synthetic samples than full dataset augmentation.

While class imbalance represents a significant factor affecting model performance, the effectiveness of regional augmentation varies across different data contexts. Note that applying SMOTE to the entire dataset generally produced metric improvements but at the cost of substantially expanding the training corpus with synthetic samples. This increases the risk of overfitting to artificial patterns. In contrast, the strategic application of SMOTE exclusively to critical regions demonstrated an advantageous balance between performance enhancement and preservation of the original data structure. This selective approach addresses slice-specific performance bottlenecks while minimizing synthetic data, preserving the dataset's statistical properties, and supporting generalization to real-world distributions.

Building on these positive results, we expanded our analysis to better understand problematic regions' impact. This included experiments where we systematically removed certain regions and conducted case studies to measure their impact on model performance. We also explored automatically detected problematic slices as a more refined tool for feature selection to take advantage of the insights from these problematic regions.

4.3. Results with region removal

As described in Section 3.3.2, removing problematic regions may appear to improve model performance, but it is not a viable long-term strategy. Table 5 presents results comparing models trained on the full dataset, with problematic regions removed and random regions of equivalent size removed.

Table 5. Performance comparison between models with the full dataset, removing problematic regions, and removing random regions of the same size.

Set/Region	Instances	Accuracy	Recall	F1 Score
Bank				
Bank (No SMOTE)	36,168	0.79	0.69	0.73
poutcome = success (remove)	34,963	0.77	0.66	0.69
month = October (remove)	35,576	0.79	0.67	0.70
month = November (remove)	35,717	0.78	0.68	0.72
Bank (random)	34,963	0.79	0.71	0.74
Bank (random)	35,576	0.79	0.69	0.73
Bank (random)	35,717	0.78	0.69	0.72
Census				
Census (No SMOTE)	30,162	0.88	0.83	0.85
Relationship = 5 $\wedge$ CG = 0 (remove)	28,932	0.86	0.79	0.82
Edu = 14 $\wedge$ CG = 0 (remove)	29,795	0.91	0.87	0.89
MS = 2 (remove)	16,097	0.88	0.62	0.64
Census (random)	28,932	0.88	0.82	0.85
Census (random)	29,759	0.90	0.85	0.87
Census (random)	16,097	0.86	0.82	0.84

In some cases, performance improved after region removal, suggesting that the excluded area contained noise or inconsistent patterns that hindered generalization. In other cases, performance declined, indicating that the removed region held critical information necessary for effective generalization. Finally, there were cases where removal had minimal impact, implying that the region was not particularly representative or influential in shaping overall performance.

One illustrative case involves a region initially flagged as problematic but that, upon closer inspection, plays a key role in supporting minority class representation. For instance, regions such as those defined by `marital status` or `education level` in the Census dataset showed this behavior. These regions comprise a substantial portion of the dataset and, while slightly imbalanced, maintain a much more balanced class distribution compared to the rest of the data, which concentrates the bulk of the overall imbalance. Removing such regions led to a notable drop in performance, especially for the minority class, with recall and F1-score substantially reduced. In contrast, removing a random region of equivalent size had little impact, as it preserved the overall class proportions more effectively. The observed performance drop can be attributed to the disproportionate removal of minority class samples, which shifts the model's focus toward the majority class. By comparison, random removal avoids this issue, enabling the model to maintain its generalization ability. These findings indicate that some regions may provide valuable class balance and support model learning despite being labeled as problematic. As a result, their removal can hinder generalization rather than improve it.

4.4. Case study

When analyzing the problem regions, we identified another recurring pattern in addition to imbalance: continuous features with highly dispersed values. In some cases, these features contained hundreds of distinct values—similar to near-unique identifiers—which led to minimal overlap between samples and limited the model's ability to extract generalizable patterns. Although `Slice Finder` automatically applies binning to features with high cardinality—specifically, those with more than half of the unique values—we observed that excessive cardinality can still degrade the region's quality. Following the pipeline proposed in section 3.3.3, we discretized these continuous features using quantile-based binning to address this issue, guided by the analysis of the univariate distribution, the number of unique values, and the proportion of these values compared to the entire set. This transformation increased the sampling density within each binning, which resulted in improved model learning and enabled the effective use of balancing methods such as SMOTE. In this sense, `Slice Finder` serves as the primary tool for identifying candidate features, and our core approach—balancing with SMOTE in combination with quantile-based binning—enables more robust treatment of problematic regions.

We focus our analysis on the Census dataset. We first identified problematic regions across the entire dataset and recorded the frequency of each attribute involved. This revealed a set of recurring attributes and their frequencies, including *Age* (85), *HoursPerWeek* (54), *Education* (41), *Education-Num* (41), *CapitalLoss* (35), and *Occupation* (34), among others. The attributes *Age* and *HoursPerWeek*, in particular, exhibited a common issue: continuous values with a wide range and sparse distribution across samples, leading to fragmented subgroups. To address this, we applied discretization by binning these continuous features into intervals of similar frequency distribution. During this process, we observed that *Education* and *Education-Num* conveyed equivalent information, with nearly identical patterns in problematic regions and visualizations. Consequently, we removed *Education-Num* from the dataset to eliminate redundancy. Following this refinement, we conducted a new round of problematic region detection, this time using the reshaped dataset. The analysis revealed new and more diverse regions, often characterized by a higher density of examples. Among the most frequently occurring attributes were: *Country* (24), *Education* (23), *CapitalGain* (14), *Age_Binned* (14), *CapitalLoss* (10), and *HoursPerWeek_Binned* (10), among others. With these updated results, we revisited our initial hypothesis and examined the class imbalance within the newly identified regions. In many cases, the imbalance persisted. To evaluate the effectiveness of targeted intervention, we applied SMOTE selectively to a subset of these regions. The results of this procedure are presented in Table 6.

Table 6. Effect of selective interventions on problematic regions of Census.

Set/Region	Instances	Accuracy	Recall	F1 Score	Log Loss	Time (s)
Census						
Census (No SMOTE)	30,162	0.88	0.83	0.85	0.23	352,70
Census + SMOTE	45,308	0.91	0.93	0.92	0.16	617,38
Edu = 14 ∧ Sex = 1 + SMOTE	30,435	0.90	0.85	0.87	0.22	413,66
Edu = 14 + SMOTE	30,432	0.90	0.85	0.87	0.22	359,88
MS = 4 + SMOTE	38,948	0.91	0.91	0.91	0.18	667,02
MS = 4 ∧ MS = 5 + SMOTE	39,755	0.90	0.90	0.90	0.21	704.42
hoursperweek_binned = 2 + SMOTE	35,110	0.90	0.91	0.91	0.20	371,00

We observed that the problematic regions identified tend to concentrate around specific features. Even in cases where the number of added synthetic samples was relatively small, such as in the region defined by *Edu = 14 + SMOTE*, performance metrics surpassed both the baseline model and those regions intervened upon solely based on the initial discovery process. Among the regions examined, those involving the attributes *HoursPerWeek_Binned* and *MaritalStatus* were particularly relevant due to high imbalance and the size of the affected subsets. For *MaritalStatus*, the region defined by value 2 (interpreted as Married-AP-spouse) was especially relevant, concentrating nearly 80% of all positive class instances. Rather than applying oversampling directly to this dominant region, we focused on complementary subregions—specifically those corresponding to values 4 and 5—where imbalance remained high, and the potential for performance gains was greater. For the *HoursPerWeek_Binned* feature, interval 2 emerged as a key problematic region. After confirming the significant class imbalance in this subgroup, we applied SMOTE selectively, using a partial oversampling rate rather than enforcing a full class balance. The outcomes showed that this partial strategy yielded performance metrics comparable to full SMOTE while reducing the number of synthetic samples and computational costs. This configuration improved recall and reduced log loss compared to the automatic balancing approach, indicating an effective increase in model sensitivity to the minority class without degrading the performance of the majority class. These results suggest that a selective and context-aware application of oversampling can yield substantial improvements while minimizing the risks of overfitting.

5. Related Work

Existing techniques for improving machine learning model performance have focused primarily on hyperparameter tuning. However, data-centric approaches are increasingly gaining attention. These include methods such as cleaning and enriching training data [Polyzotis et al. 2017], using structured explanation formats [El Gebaly et al. 2014], identifying data subsets for training more effective models [Ribeiro et al. 2023]. This paper discusses practical strategies to address these problematic regions, aiming to improve model performance by mitigating the locally observed class imbalance. Other approaches, such as bias correction techniques, have been explored in previous studies through preprocessing interventions [Lin et al. 2024, Kamiran and Calders 2012].

Among the algorithms designed to identify problematic regions, the following stand out: `Slice Finder` [Chung et al. 2020] identifies statistically significant subpopulations (slices) where the model underperforms, using metrics such as effect size and log loss. It systematically searches for feature-value combinations to find interpretable groups with poor metrics; `SliceLine` [Sagadeeva and Boehm 2021] seeks to find all possible problematic regions in the set using an iterative process based on model errors with vectorized linear algebra, enumerating them and pruning them with sparse linear algebra; `DivExplorer` [Pastor et al. 2021] identifies discrepancies in the behavior of machine learning models across different regions of data. The algorithm bases its definition on the divergence of the error distribution, allowing it to detect and analyze areas where the model performs inconsistently. Despite their relevant contribution, these approaches do not provide practical guidance on using the information obtained to generate insights into our models and improve their performance, which is left to the user. Our work addresses this issue through interventions guided by problematic regions.

6. Conclusions

This study demonstrates that targeted interventions in problematic regions can improve model performance. The impact of each intervention depends on factors such as sample density and the range of feature values. Using simple analytics and local balancing strategies, we observed more substantial improvements compared to global methods. Our approach identifies regions of poor performance and actively uses this information to guide interventions like SMOTE, which are applied only where needed. This leads to better resource use, reduced risk of overfitting, and improved evaluation metrics. Future work includes refining how numerical features are handled, exploring ways to merge related regions, and integrating corrective steps directly into the region discovery process—so that the output includes both the problematic slices and suggestions for data improvement.

7. Artifacts

The source code and datasets for our experimental evaluation are available in our public repository: `https://github.com/GregullyWillian/SF-Study.git`.

8. Acknowledgments

This research was partially funded by the Araucária Foundation for Scientific and Technological Development of Paraná.

References

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*, 16:321–357.

Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., and Whang, S. E. (2020). Automated data slicing for model validation: A big data - ai integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2284–2296.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.

El Gebaly, K., Agrawal, P., Golab, L., Korn, F., and Srivastava, D. (2014). Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment*, 8(1):61–72.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2):429–444.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kerrigan, D. and Bertini, E. (2023). Slicelens: Guided exploration of machine learning datasets. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–7.

Kohavi, R. (1996). Census Income. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP7S.

Lin, Y., Gupta, S., and Jagadish, H. (2024). Mitigating subgroup unfairness in machine learning classifiers: A data-driven approach. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 2151–2163. IEEE.

Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, USA.

Moro, S., Rita, P., and Cortez, P. (2012). Bank Marketing. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

Pastor, E., De Alfaro, L., and Baralis, E. (2021). Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1400–1412.

Polyzotis, N., Roy, S., Whang, S. E., and Zinkevich, M. (2017). Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1723–1726.

Ribeiro, V., Pena, E. H. M., Saldanha, R., Akbarinia, R., Valduriez, P., Khan, F., Stoyanovich, J., and Porto, F. (2023). Subset modelling: A domain partitioning strategy for data-efficient machine-learning. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 318–323, Porto Alegre, RS, Brasil. SBC.

Sagadeeva, S. and Boehm, M. (2021). Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2290–2299.

Sakar, C. and Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F88Q.

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

Yang, L. and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316.

Zhang, X., Ono, J. P., Song, H., Gou, L., Ma, K.-L., and Ren, L. (2023). Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852.