1 Introduction

In recent years, the field of machine learning has evolved from an emerging science into a widely applied technology, finding applications across various domains such as business, industry, and scientific research. However, as machine learning techniques have gained prominence, a critical challenge has surfaced known as the class imbalance problem. This problem, characterized by a significant disparity in the number of samples between different classes within a dataset, has profound implications for classification performance and decision-making in real-world applications.

It has become increasingly apparent that imbalanced datasets can lead to suboptimal classification results, prompting researchers to explore solutions to mitigate its impact. The class imbalance problem is widespread, affecting a substantial portion of the data mining community. It is essential to understand that class imbalances can manifest in various application domains, including fraud detection, risk management, text classification, and medical diagnosis. In some cases, these imbalances are inherent to the problem, while in others, they arise due to limitations in data collection processes or the need for human intervention in selecting examples for training.

Both data-level and algorithmic-level solutions have been proposed to address class imbalance. These include resampling techniques such as oversampling and undersampling, adjustments to class-specific costs, threshold tuning, and recognition-based learning approaches. Researchers have dedicated considerable effort to developing and refining these methods, aiming to improve the performance and fairness of machine learning models in the face of imbalanced data.

In this report, we illustrate a comparative analysis aiming to highlight the influence of data balancing and other factors in the process of classifying imbalanced datasets. In addition to assessing the performance of established data-level balancing algorithms, our objective includes the exploration of diverse classification scenarios wherein various factors may exert an influence on classification efficacy.

The report is organized into distinct sections, including Methods, Results, and Discussion. Within the Methods section, we comprehensively elucidate our approach, which comprises a four-step pipeline encompassing data generation, balancing, classification, and output analysis. It's noteworthy to emphasize that each step of this methodology is supported by extensive literature research, ensuring a robust foundation for our approach.

In the Results section, we provide the most significant findings derived from the systematic variation of parameters within our defined pipeline. Through rigorous experimentation, we explore the impact of parameter adjustments on the classification outcomes. To facilitate a comprehensive understanding of our findings, we employ data visualization techniques and provide insightful plots that vividly represent the observed results.

In the final section, the Discussion, we delve into a thorough analysis of the results and draw meaningful conclusions from our study. We closely examine what our findings mean in practical terms and how they can be applied in real-world situations. Additionally, we consider areas where our study could be improved or enhanced for future research. By acknowledging both the strengths and limitations of our work, we lay the groundwork for future studies to build upon our findings and advance the field of classifying imbalanced datasets.

2 Methods

2.1 Pipeline description

2.2 Data balancing

As for the data balancing step, here are some important characteristics that our method should focus on:

- The balancing should be generalizable and applicable to different levels of imbalancement and subsequent classification algorithms;

- It should mantain the actual class structure and represent accurately the minority class pattern;
- It should support a correct class identification.

We hereby summarize the main features of the data-level balancing algorithms that we implemented in the pipeline. Typically, these methods are categorized into three main groups: synthetic samplers, resamplrs and hybrid samplers.

SMOTE (Synthetic Minority Oversampling Technique) is probably the most known of the balancing methods. SMOTE is an over-sampling technique designed to address class imbalance. It generates synthetic minority class samples by considering feature vectors. For each minority class sample, it calculates the differences between the sample and its nearest neighbors. A random scaling factor between 0 and 1 is applied to these differences, and the results are added to the original sample's feature vector. This process creates new data points along line segments between features, effectively making the decision boundary of the minority class more inclusive. This approach helps rebalance class distribution and improves model performance in imbalanced datasets.

3 Evaluating a New Marker for Risk Prediction Using the Test Tradeoff

3.1 Definitions (adapted to previous notation)

Let ρ be a risk score a model outputs, where $\rho(X)$ is the score for instance X. Let $D \in \{0, 1\}$ be the r.v. that states disease $D = 1$ or not disease $D = 0$. Then probability of developing disease if the risk score is $\rho(X) = s$ is given by

$$R_s = \mathbb{P}(D = 1|\rho(X) = s)$$

Requirement: If $s_1 < s_2 \Rightarrow R_{s_1} < R_{s_2}$ Cutoff τ defined such that

$$\rho(X) \geq \tau \Rightarrow \hat{Y} = 1$$
$$\rho(X) < \tau \Rightarrow \hat{Y} = 0$$

Then the True Positive Rate and the False Positive Rate are

$$TPR(\tau) = \mathbb{P}(\rho(X) \geq \tau|D = 1)$$
$$FPR(\tau) = \mathbb{P}(\rho(X) \geq \tau|D = 0)$$

We say the probability of having/developing the disease is $\mathbb{P}(D = 1)$ Let T be treatment or no treatment and define the utilities

$U_{T,D}$	utility of treating person with disease
$U_{\neg T,D}$	utility of not treating person with disease
$U_{\neg T,\neg D}$	utility of not treating person without disease
$U_{T,\neg D}$	utility of treating person without disease

and U_{test} the negative utility / harm of a test.

3.2 Assumptions

Assumptions:

$$U_{T,D} > U_{\neg T,D}$$
$$U_{\neg T,\neg D} > U_{T,\neg D}$$
$$U_{test} < 0$$

3.3 Comparison of the Maximum Expected Utility of Risk Prediction

Comparison of the net benefit of Model 2 versus Model 1 is a comparison of the maximum expected utility of risk prediction under Model 2 versus Model 1. The maximum expected utility of risk prediction is the maximum, over the cutpoints, of the expected utility of risk prediction. The **expected utility of risk prediction** is

$$\begin{aligned} U_{pred}(\tau) =& P\mathbb{P}(\rho(X) \geq \tau | D = 1)U_{T,D} \\ &+ P\mathbb{P}(\rho(X) < \tau | D = 1)U_{\neg T,D} \\ &+ (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)U_{T,\neg D} \\ &+ (1-P)\mathbb{P}(\rho(X) < \tau | D = 0)U_{\neg T,\neg D} \\ &+ U_{test} \end{aligned}$$

These definitions allow to express **Fundamental Rule Version 1:**

$$\text{select model 2 if}: \quad \max_{\tau} U^2_{pred}(\tau) > \max_{\tau} U^1_{pred}(\tau)$$

Comparing maximum expected utilities of risk prediction: a simplification involving the no treatment option

Setting the probabilities for no treatment to 1 and the ones for treatment to 0 the expected utility of no treatment at all is:

$$U_{\neg T} = PU_{\neg T,D} + (1-P)U_{\neg T,\neg D}$$

and analogously the expected utility of treating everyone is

$$U_T = PU_{T,D} + (1-P)U_{T,\neg D}$$

The expected utility of risk prediction in excess of the expected utility of no treatment is thus

$$\begin{aligned} U_{pred*}(\tau) =& U_{pred}(\tau) - U_{\neg T} = \\ &+ P\mathbb{P}(\rho(X) \geq \tau | D = 1)U_{T,D} \\ &+ P\mathbb{P}(\rho(X) < \tau | D = 1)U_{\neg T,D} \\ &+ (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)U_{T,\neg D} \\ &+ (1-P)\mathbb{P}(\rho(X) < \tau | D = 0)U_{\neg T,\neg D} \\ &+ U_{test} \\ &- PU_{\neg T,D} - (1-P)U_{\neg T,\neg D} \\ =& P\mathbb{P}(\rho(X) \geq \tau | D = 1)U_{T,D} \\ &- P\mathbb{P}(\rho(X) \geq \tau | D = 1)U_{\neg T,D} \\ &+ (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)U_{T,\neg D} \\ &- (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)U_{\neg T,\neg D} \\ &+ U_{test} \\ =& P\mathbb{P}(\rho(X) \geq \tau | D = 1)(U_{T,D} - U_{\neg T,D}) \\ &+ (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)(U_{T,\neg D} - U_{\neg T,\neg D}) \\ &+ U_{test} \end{aligned} \qquad P\mathbb{P}(\rho(X) \geq \tau | D = 1)U_{T,D}$$

When defining

$$\begin{aligned} B :=& U_{T,D} - U_{\neg T,D} \\ C :=& U_{T,\neg D} - U_{\neg T,\neg D} \end{aligned}$$

we can write this as

$$U_{pred*}(\tau) = P\mathbb{P}(\rho(X) \geq \tau | D = 1)B + (1-P)\mathbb{P}(\rho(X) \geq \tau | D = 0)C + U_{test}$$

The **risk threshold**, denoted T, is the probability of developing disease in the population at which the expected utility of treatment and no treatment is the same. We obtain it by substituting T for P when setting $U_{\neg T} = U_T$ in equations and solving for T to get the following formula for the risk threshold,

$$T = \frac{C}{C+B}$$

and for the odds of the risk threshold this gives

$$\frac{T}{1-T} = \frac{C}{B}$$

3.4 Net Benefit

The **Net Benefit** $NB(\tau)$ is defined as

$$\begin{aligned}
NB(\tau) =& \frac{U_{pred}(\tau) - U_{\neg T}}{B} \\
=& \frac{P\mathbb{P}(\rho(X) \geq \tau | D=1)B + (1-P)\mathbb{P}(\rho(X) \geq \tau | D=0)C + U_{test}}{B} \\
=& P\mathbb{P}(\rho(X) \geq \tau | D=1) + (1-P)\mathbb{P}(\rho(X) \geq \tau | D=0)\frac{C}{B} + \frac{U_{test}}{B} \\
& \text{substituting the risk threshold} \\
=& P\mathbb{P}(\rho(X) \geq \tau | D=1) + (1-P)\mathbb{P}(\rho(X) \geq \tau | D=0)\frac{T}{1-T} + \frac{U_{test}}{B}
\end{aligned}$$

We can also see that, with $P = \mathbb{P}(D=1)$ we can replace the conditional with a joint probability

$$\begin{aligned}
NB(\tau) =& P\mathbb{P}(\rho(X) \geq \tau | D=1) + (1-P)\mathbb{P}(\rho(X) \geq \tau | D=0)\frac{T}{1-T} + \frac{U_{test}}{B} \\
=& \mathbb{P}(\rho(X) \geq \tau, D=1) + \mathbb{P}(\rho(X) \geq \tau, D=0)\frac{T}{1-T} + \frac{U_{test}}{B} \\
& \text{for concrete model the joint probabilities represent the TP and FP respectively} \\
=& \frac{\text{TP}}{N} + \frac{T}{1-T}\frac{\text{FP}}{N} + \frac{U_{test}}{B}
\end{aligned}$$

So in absence of testing costs we have:

$$NB(\tau) = \frac{\text{TP} + \frac{T}{1-T}\text{FP}}{N}$$

This net benefit for a decision curve, NB, is the maximum benefit of risk prediction (in excess of the benefit of no treatment) in units of the benefit of treating a true positive. It equals the benefit of treating a true positive after subtracting the cost of treating a false positive at an "exchange rate" based on the risk threshold.

Optimization Requirement, $R_t = T$, $ROCSLOPEt = \frac{1-P}{P}\frac{T}{1-T}$

4 The harm of class imbalance corrections for risk prediction models

The paper argues that the class imbalance is not a pervasive problem for prediction model development. First, the problem is specific to the classification accuracy measure. The limitations of focusing on classification accuracy as a measure of predictive performance is well known. Second, if we consider models that produce estimated probabilities of the event of interest, an adjustment of the classification threshold probability can be used to ensure adequate classification performance (ie, probability threshold to classify individuals as high risk does not have to be 0.5). A probability threshold to select individuals for a given treatment implies certain misclassification costs and should be determined

using clinical considerations. If we use a probability threshold of 0.1 to classify individuals as high risk and suggest a specific treatment, this means that we accept to treat up to 10 individuals in order to treat 1 individual with the event: we accept up to 9 false positives, or unnecessary treatments, per true positive.

4.1 Discussion

¡¡¡¡¡¡¡ HEAD ======= ¿¿¿¿¿¿¿ baseline The key finding of our work is that training logistic regres-

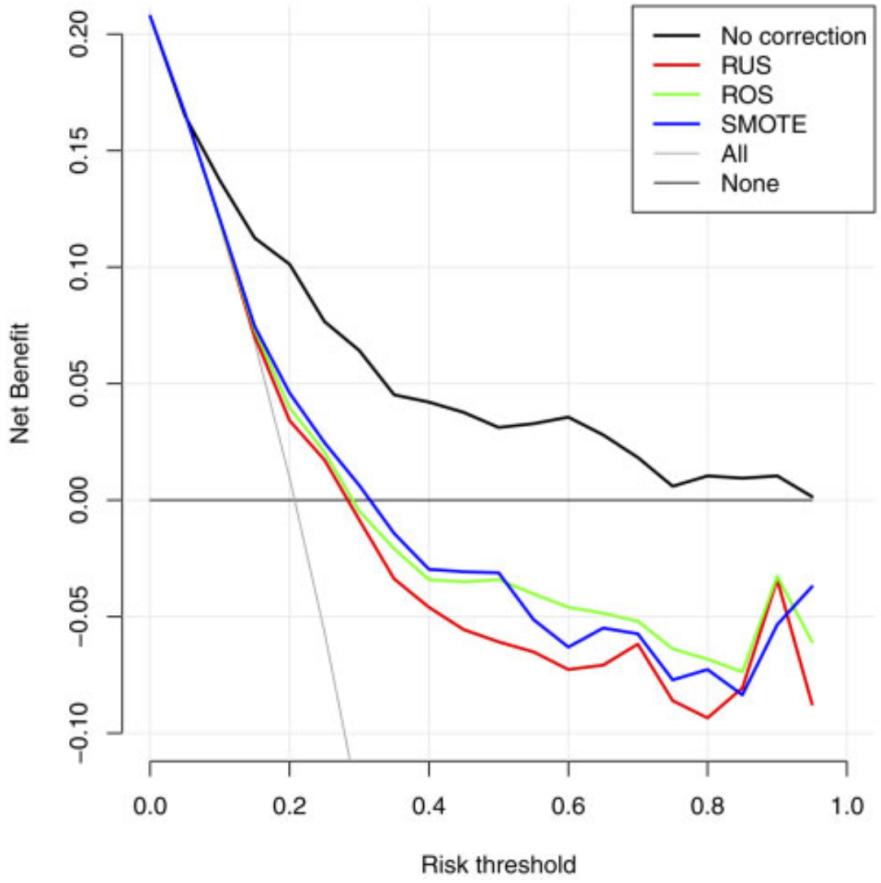

Figure 1: Risk vs Net Benefit for uncorrected vs ROS, RUS and SMOTE

sion models on imbalance corrected data did not lead to better AUROC compared to models trained on uncorrected data, but did result in strong and systematic overestimation of the probability for the minority class. In addition, all imbalance corrections had negative consequences for the calibration slope. The lower the event fraction, the more outspoken the results.

Strong miscalibration reduces the clinical utility of a prediction model.30 Models yielding probability estimates that are clearly too high may lead to overtreatment. For example, if a model overestimates the risk of malignancy of a detected ovarian tumor, the decision to refer patients to specialized surgery may be taken too quickly. Class imbalance is often framed as problematic in the context of prediction models that classify patients into low-risk versus high-risk groups. Nevertheless, for clinical prediction models the accurate estimation of probabilities is essential to help in defining such low-risk and high-risk groups. For instance, clinical staff using the model to support treatment decisions may choose probability thresholds ¡¡¡¡¡¡¡ HEAD to match the assumed misclassification costs that best fit the context.

In contrast, our study suggests that, at least for logistic regression models, RUS (or ROS or SMOTE) is unlikely to lead to better discrimination or separability between the minority and majority

classes.

4.2 Conclusion

Our study shows that correcting class imbalance did not result in better prediction models based on standard or ridge logistic regression. The imbalance corrections resulted in inaccurate probability estimates without improving discrimination in terms of AUROC.

======= to match the assumed misclassification costs that best fit the context. ¿¿¿¿¿¿¿ baseline

5 Metrics

¡¡¡¡¡¡¡ HEAD

5.1 Accuracy rate

If benefit and harm are weighted equally, the odds of the threshold is 1:1, or a threshold probability of 50%. This cutoff is by default considered in the calculation of the error rate, which is defined as (FN + FP)/N. The complement is the accuracy rate: (TN + TP)/N. Often FN classifications are more important than FP classifications, which makes the accuracy rate not a sensible indicator of clinical usefulness. Other disadvantages include that the accuracy rate by definition is high for a frequent or infrequent outcome (i.e. imbalanced data).

The accuracy rate is usually calculated at the simplistic cutoff of 50%, but can also be calculated at clinically defendable thresholds. The harm-to-benefit ratio that underpins the choice of the cutoff should then be used to calculate a weighted accuracy, or its complement, the weighted error rate. We can express the TN classifications in units of the TP classifications, such that the weighted accuracy is calculated as (TP + w TN)/n.

The improvement that is obtained by making decisions based on predictions from the model is the difference between the weighted accuracy obtained with the model versus the weighted accuracy of the default policy.

=======

¿¿¿¿¿¿¿ baseline

5.2 Receiver Operator Characteristic (ROC)

Consider a binary classification problem where we have a classifier that outputs a continuous score s(X) for an instance X. We then set a threshold τ to decide the predicted class. If $s(X) > \tau$, we predict the positive class; otherwise, we predict the negative class.

$$TPR(\tau) = \mathbb{P}(s(X) > \tau | Y = 1)$$
$$FPR(\tau) = \mathbb{P}(s(X) > \tau | Y = 0)$$

Here, Y is the true class of an instance. TPR is the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance. Similarly, FPR is the probability that the classifier ranks a randomly chosen negative instance higher than a randomly chosen positive instance.

The ROC curve plots $TPR(\tau)$ against $FPR(\tau)$ for all possible thresholds τ, producing a curve that ranges from $(0, 0)$ to $(1, 1)$. We can interpret a ROC plot as plotting the path of a function

$$f : \mathbb{R} \to [0, 1]^2, \quad \tau \mapsto (TPR(\tau), FPR(\tau))$$

The Area Under the ROC Curve (AUC) then provides a single scalar value that represents the expected performance of the classifier. An AUC of 1 indicates a perfect classifier, while an AUC of 0.5 indicates a classifier that performs no better than random chance.

AUC can also be interpreted in terms of the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, assuming that one positive and one negative instance are chosen at random. ¡¡¡¡¡¡¡ HEAD The area under the curve (AUC) can be interpreted as the probability that a patient with the outcome is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome Some may consider the interpretation of AUC as straightforward. Others may object that we consider a pair of subjects, one with and one without the outcome, and that such conditioning is a rather artificial situation. Statistically, this conditioning on a pair of patients is attractive, since it makes the area independent of the incidence of the outcome (or event rate), in contrast to R2 or the Brier score, for example. The AUC is a rank order statistic. As a rank order statistic, it is insensitive to errors in calibration such as differences in average outcome. Confidence intervals for the AUC (or c statistic) can be calculated with various methods. Standard asymptotic methods may be problematic, especially when sensitivity or specificity is close to 0% or 100%. Bootstrap resampling is a good choice for many situations.

5.3 Net Benefit

The choice of risk threshold implicitly conveys the adopted relative misclassification costs. It can be derived that the odds of the risk threshold equal the harm-to-benefit ratio, which is the harm of a false positive divided by the benefit of a true positive. For example, if a risk threshold of 20% is used, the odds are 1 to 4. Therefore, a 20% risk threshold assumes that the harm of a false positive is one-quarter of the benefit of a true positive or that 1 true positive is worth 4 false positives: A clinician might express this in terms such as "I would not do more than five biopsies to find one cancer." Hence, when applying a model to a set of patients, we can correct the number of true positives (TP) for the number of false positive (FP) using the odds w of the risk threshold t:

$$\text{TP} - w\text{FP} = \text{TP} - \frac{\tau}{1-\tau}\text{FP}$$

When dividing by the total sample size N, the Net Benefit is obtained

$$\text{NB} = \frac{1}{N}(\text{TP} - w\text{FP}) = \frac{1}{N}(\text{TP} - \frac{\tau}{1-\tau}\text{FP})$$

The Net Benefit of treat-none is always 0, whereas the Net Benefit of treat-all is positive for risk thresholds below the event rate and negative for risk thresholds above the event rate.

5.4 Net Benefit and Risk Threshold

Suppose r.v. Y with outcome diseased D or healthy $\neg D$. Let $c_{TP}, c_{FP}, c_{TN}, c_{FN}$ then expected cost of predicting $\neg D$ is

$$\mathbb{P}(Y = 1)c_{FN} + \mathbb{P}(Y = 0)c_{TN}$$

and expected cost for predicting D is

$$\mathbb{P}(Y = 1)c_{TP} + \mathbb{P}(Y = 0)c_{FP}$$

If we write $T = \mathbb{P}(Y = 1)$ we can rewrite to

$$Tc_{FN} + (1 - T)c_{TN}$$

and

$$Tc_{TP} + (1 - T)c_{FP}$$

One is indifferent about treatment if both expected costs are equal.

$$\begin{aligned} Tc_{FN} + (1-T)c_{TN} &= Tc_{TP} + (1-T)c_{FP} \\ &\Leftrightarrow \\ T &= \frac{c_{TN} - c_{FP}}{(c_{TN} - c_{FP}) + (c_{TP} - c_{FN})} \end{aligned}$$

It can also be rearranged in a different way

$$\begin{aligned} Tc_{FN} + (1-T)c_{TN} &= Tc_{TP} + (1-T)c_{FP} \\ &\Leftrightarrow \\ \frac{T}{1-T} &= \frac{c_{TN} - c_{FP}}{c_{TP} - c_{FN}} \end{aligned}$$

In the first group, the costs relate to undertreatment (false-negative classifications) The costs of these false-negative classifications c_{FN} should be compared to the costs of true-positive classifications c_{TP}. The difference $c_{TP} - c_{FN}$ is the net benefit of treating all who have the disease compared to treating none of them. Suppose you have a treatment that causes lots of damage as well as curing the disease. Then this difference might be low, i.e. treating everyone with the dangerous treatment does not give much better outcome than simply not treating anyone. Suppose on the other hand you have a devastating disease like Polio and a low cost treatment like a Polio-vaccine, then that difference will be strongly positive.

In the second group, relevant costs are for those without the event if not treated, who are treated ("overtreated"). The costs of these false-positive classifications (c_{FP}) should be compared to the costs of true-negative classifications (c_{TN}) while $c_{TN} - c_{FP}$ is the harm of treating all who don't have the disease compared to the benefit of not treating any of them. E.g. the cost of not treating anyone without the disease might be 0 but the cost of treating them might be high. Then this value is strongly negative. On the other hand suppose again a low impact vaccine that barely does harm, then that difference may be small negative.

Odds (cutoff) = Harm/Benefit

The ratio in case of e.g. the polio vaccine could be strongly in favour of treating everyone (small cost for those without, high benefit for those with disease).

5.5 Calibration

Another key property of a prediction model is calibration, i.e., the agreement between observed outcomes and predictions.

$$TPR(\tau) = \mathbb{P}(s(X) > \tau | Y = 1)$$

5.6 Mixed

$$\begin{aligned} TPR(\tau) &= \mathbb{P}(s(X) > \tau | Y = 1) \\ FPR(\tau) &= \mathbb{P}(s(X) > \tau | Y = 0) \end{aligned}$$

=======
¿¿¿¿¿¿¿ baseline