1 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.

1.1 Discussion

The key finding of our work is that training logistic regression 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 to match the assumed misclassification costs that best fit the context.

2 Metrics

2.1 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.

3 Metrics

3.1 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.

3.2 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.

3.2.1 Discussion

The key finding of our work is that training logistic regression 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 to match the assumed misclassification costs that best fit the context.