Proposal: assessing the impact of class imbalance corrections on model calibration.

Alex Carriero (9028757)

Date: 13/10/2022

Word Count: 735

Program: Methodology and Statistics for Behavioral, Biomedical, and Social Sciences.

Supervisors: Maarten van Smeden (Utrecht Medical Center Utrecht), Ben van Calster (Leuven University and Leiden University Medical Center) and Kim Luijken (Utrecht Medical Center Utrecht).

Host Institution: Julius Center for Health Science and Primary Care, UMC.

Candidate Journal: Statistics in Medicine.

FETC-approved: 22-1809

1 | Introduction

Prediction modelling in medicine is gaining increasing attention; clinicians are often interested in predicting a patient's risk of disease. Due to the (thankfully) rare nature of many diseases, the data available to train clinical prediction models are often heavily imbalanced (i.e., the number of patients in one class dramatically outnumbers the other) [1]. This is referred to as class imbalance. Class imbalance is seen major problem in the field of machine learning as it is known to degrade model performance [2]. Consequently, imbalance correction methodologies are proposed as a solution [2].

An ideal imbalance correction will improve all aspects of model performance. These criteria include: classification accuracy, discrimination and calibration. Accuracy refers to the proportion of patients that a model classifies correctly (after a risk threshold is imposed). Discrimination refers to a model's ability to yield higher risk estimates for patients in the positive class than for those in the negative class. Finally, calibration refers to the reliability of the risk predictions themselves; for instance, a poorly calibrated model may produce risk predictions that consistently over- or under-estimate reality, or produce risk estimates which are too extreme (too close of 0 or 1) or too modest [3].

Class imbalance is not unique to medical data sets and literature introducing imbalance correction methods arises from many disciplines. An abundance of imbalance corrections exist and are summarized by [4, 5, 1, 6, 7]. Information regarding the effect of these corrections on model calibration is sparse. In medicine, it is essential that model calibration is assessed. This is because in practice, risk predictions from the model are given directly to a clinician who will use the information to council patients and inform treatment decisions. If a model is poorly calibrated, the personal costs to the patient may be enormous. It is entirely possible for a model to exhibit excellent classification accuracy and discrimination while calibration is poor [3]. Therefore, assessing only discrimination and accuracy is insufficient.

Only one study has assessed the impact of imbalance corrections on model calibration. Goorbergh et al. [8] demonstrated that implementing imbalance corrections lead to dramatically deteriorated model calibration, to the extent that no correction was recommended [8]. In this study, models were developed using logistic regression and penalized (ridge) logistic regression [8]. Motivated by the work of Goorbergh et al. [8], we must ensure that the "cure" is not worse than the disease. In our research, we aim to assess the impact of imbalance corrections on model calibration for prediction models trained with a wider variety of classification algorithms including: linear classifiers, ensemble learning algorithms and algorithms specifically designed to handle class imbalance. Furthermore, we aim to answer the question: can imbalance corrections improve overall model performance without comprising model calibration?

2| Analytic Strategy

We will evaluate the performance of several imbalance corrections in a simulation study. We will adhere to the ADEMP guidelines for the design and reporting of our simulation [9].

2.1| Simulation Study

The aim of the simulation study is to determine which pair(s) of imbalance correction and classification algorithm can outperform the classification algorithms without imbalance corrections.

Imbalanced data will be simulated to reflect 27 scenarios. The following criteria will be varied: number of predictors, event fraction and sample size. The number of predictors will vary through the set {8,16,32} and event fraction, through the set {0.5, 0.2, 0.02}. The minimum sample size for the prediction model (N) will be computed according to formulae from Riley et al. [10]. Sample size will then vary through the set {$\frac{1}{2}$N, N and 2N}.

Under each scenario, 2000 data sets will be generated. More specifically, test and training data will be generated such that the training set is 10x larger than the test set. Each simulated data set will be analysed by 30 methods = 6 (classification algorithms) x 5 (imbalance corrections). The classification algorithms and imbalance corrections we will include in our simulation are detailed in Table 1.

Table 1: Classification algorithms and imbalance corrections to be evaluated.

Index	Classification Algorithms	Imbalance Corrections
1	Logistic Regression	None
2	Support Vector Machine	RUS (random under sampling)
3	Random Forest	ROS (random over sampling)
4	XG Boost	SMOTE (synthetic majority over sampling)
5	RUSBoost	SMOTE - ENN (SMOTE - edited nearest neighbours)
6	Easy Ensemble	

Finally, performance criteria will include measures of model discrimination, accuracy and calibration. Discrimination will be measured by area under the receiver operator curve (AUROC). Classification accuracy by Matthew's correlation coefficient (MCC), overall accuracy, sensitivity and specificity. Finally calibration will be measured in terms of calibration intercept and calibration slope.

2.2| Software

All analyses will be conducted using the open source statistical software R [11]. Additionally, our simulation study is expected to be quite computationally intensive. Therefore, we intend to run the simulation using the high performance computers at the UMC.

References

[1] Satyam Maheshwari, R.C. Jain, and R.S. Jadon. "An Insight into Rare Class Problem: Analysis and Potential Solutions". In: *Journal of Computer Science* 14.6 (May 2018), pp. 777–792.

[2] Lian Yu and Nengfeng Zhou. *Survey of Imbalanced Data Methodologies*. 2021. DOI: 10.48550/ARXIV.2104.02240. URL: https://arxiv.org/abs/2104.02240.

[3] Ben Van Calster et al. "Calibration: the Achilles heel of predictive analytics". In: *BMC Medicine* 17.1 (2019), p. 230.

[4] Paula Branco, Luis Torgo, and Rita P. Ribeiro. "A Survey of Predictive Modeling on Imbalanced Domains". In: 49.2 (Aug. 2016). ISSN: 0360-0300. DOI: 10.1145/2907070. URL: https://doi.org/10.1145/2907070.

[5] Victoria López et al. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". In: *Information Sciences* 250 (2013), pp. 113–141.

[6] Guo Haixiang et al. "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73 (2017), pp. 220–239.

[7] Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232.

[8] Ruben van den Goorbergh et al. "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression". In: *Journal of the American Medical Informatics Association* 29.9 (June 2022), pp. 1525–1534. ISSN: 1527-974X. DOI: 10.1093/jamia/ocac093. URL: https://doi.org/10.1093/jamia/ocac093.

[9] Tim P. Morris, Ian R. White, and Michael J. Crowther. "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11 (2019), pp. 2074–2102.

[10] Richard D. Riley et al. "Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome". In: *Statistics in Medicine* 41.7 (2022), pp. 1280–1295.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

Further Reading

[1] Walter Bouwmeester et al. "Reporting and Methods in Clinical Prediction Research: A Systematic Review". In: *PLOS Medicine* 9.5 (May 2012), pp. 1–13.

[2] Laure Wynants et al. "Three myths about risk thresholds for prediction models". In: *BMC Medicine* 17.1 (2019), p. 192.

[3] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC Genomics* 21.1 (2020), p. 6.

[4] Victoria López et al. "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics". In: *Expert Systems with Applications* 39.7 (2012), pp. 6585–6608.

[5] Mohd Adil et al. "Solving the Problem of Class Imbalance in the Prediction of Hotel Cancelations: A Hybridized Machine Learning Approach". In: *Processes* 9.10 (2021).

[6] Prabhjot Kaur and Anjana Gosain. "Empirical Assessment of Ensemble based Approaches to Classify Imbalanced Data in Binary Classification". In: *International Journal of Advanced Computer Science and Applications* (2019).

[7] Ahmad S. Tarawneh et al. "Stop Oversampling for Class Imbalance Learning: A Review". In: *IEEE Access* 10 (2022), pp. 47643–47660. DOI: 10.1109/ACCESS.2022.3169512.

[8] Yotam Elor and Hadar Averbuch-Elor. *To SMOTE, or not to SMOTE?* 2022. DOI: 10.48550/ARXIV.2201.08528. URL: https://arxiv.org/abs/2201.08528.

[9] Maarten van Smeden et al. "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis". In: *BMC Medical Research Methodology* 16.1 (2016), p. 163.

[10] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. DOI: 10.1613/jair.953. URL: https://doi.org/10.1613%2Fjair.953.

[11] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory Undersampling for Class-Imbalance Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2009), pp. 539–550. DOI: 10.1109/TSMCB.2008.2007853.

[12] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data". In: 6.1 (June 2004), pp. 20–29. ISSN: 1931-0145. DOI: 10.1145/1007730.1007735. URL: https://doi.org/10.1145/1007730.1007735.

[13] Yüksel Özkan, Mert Demirarslan, and Aslı Suner. "Effect of data preprocessing on ensemble learning for classification in disease diagnosis". In: *Communications in Statistics - Simulation and Computation* 0.0 (2022), pp. 1–21.

[14] Chris Seiffert et al. "RUSBoost: Improving classification performance when training data is skewed". In: *2008 19th International Conference on Pattern Recognition*. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761297.

[15] Jack Wilkinson et al. "Time to reality check the promises of machine learning-powered precision medicine". In: *The Lancet Digital Health* 2.12 (2022/10/07 2020), e677–e680.

[16] Livia Faes et al. “Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?” In: *Frontiers in Digital Health* 4 (2022).

[17] Susan Athey and Guido W. Imbens. “Machine Learning Methods That Economists Should Know About”. In: *Annual Review of Economics* 11.1 (2019), pp. 685–725.

[18] Shubham Malik, Rohan Harode, and Akash Singh. *XGBoost: A Deep Dive into Boosting (Introduction Documentation)*. Feb. 2020. DOI: 10.13140/RG.2.2.15243.64803.

[19] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. “Applying Support Vector Machines to Imbalanced Datasets”. In: *Machine Learning: ECML 2004*. Ed. by Jean-François Boulicaut et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 39–50.

[20] Andrew P. Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159.

[21] Dennis L. Wilson. “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-2.3 (1972), pp. 408–421. DOI: 10.1109/TSMC.1972.4309137.

[22] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”. In: *Advances in Intelligent Computing*. Ed. by De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.

[23] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. “Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests”. In: *BMJ* 352 (2016).

[24] Ben Van Calster et al. “Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study”. In: *Statistical Methods in Medical Research* 29.11 (2020), pp. 3166–3178.

[25] “Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration”. In: *Annals of Internal Medicine* 162.1 (2015), W1–W73.