

****Proposal: Stop correcting for class imbalance?***

Alex Carriero (9028757)

Date: 13/10/2022

Word Count: XXX

Program: Methodology and Statistics for Behavioural, 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

The presence of prediction modelling in the field of medicine is rapidly increasing. In a clinical setting, the goal of prediction modelling is often to (accurately) predict a patient’s risk of experiencing an event (e.g., a stroke). In other words, to predict which binary class (event or no event) a patient is most likely to belong to. 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) [14]. This is referred to as class imbalance, and is seen as a major problem in the field of machine learning [14].

Class imbalance is thought to diminish the quality of prediction models [25]. Quality is multifaceted. It is characterized by three criteria: 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.

Calibration is the metric which is most interesting when developing clinical prediction models. This is because in practice, risk predictions from the model are given directly to a clinician who will use the information to counsel patients and inform treatment decisions. It is essential that these predictions are accurate (i.e., calibration is good). Otherwise, the personal costs to the patient may be enormous. Furthermore, it is entirely possible for a model to have great accuracy and discrimination while calibration is poor [22]. Thus, all three criteria must be considered when discussing the impact of imbalance on the quality of clinical prediction models. This is rarely the case and unfortunately, it is calibration that is often forgotten [22].

Class imbalance is not unique to medical data sets. Thus, literature focused on imbalance correction methods arises from many disciplines. An abundance of imbalance corrections exist and are well summarized by [4, 12, 14, 9, 11]. Yet, information regarding the effect of these corrections on model calibration is sparse. One recent study 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?

- expectations specified ?

2| Analytic Strategy

To provide a fair comparison of imbalance corrections a simulation study will be designed and implemented.

2.1| Simulation Study

We will adhere to the ADEMPT guidelines for the design and reporting of our simulation study [15].

The aim of the simulation study is to determine which pair(s) of imbalance correction and classification algorithm can outperform the classification algorithm alone.

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. [18]. 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.

Finally, each simulated data set will be analysed by 30 methods = 6 (classification algorithms) x 5 (imbalance corrections). Classification algorithms will include: Logistic Regression, Support Vector Machine, Random Forest, XG Boost, RUSBoost, and Easy Ensemble. Imbalance corrections will include: random under sampling, random over sampling, synthetic minority over sampling (SMOTE), SMOTE-ENN and no correction.

Finally, performance criteria will include: AUROC, MCC, accuracy, sensitivity and specificity, calibration intercept and calibration slope. Calibration intercept is the primary metric.

- something about reporting results ?

2.2| Software

All analyses will be conducted using the open source statistical software R [17]. 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] Mohd Adil et al. “Solving the Problem of Class Imbalance in the Prediction of Hotel Cancellations: A Hybridized Machine Learning Approach”. In: *Processes* 9.10 (2021).
- [2] 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.
- [3] Walter Bouwmeester et al. “Reporting and Methods in Clinical Prediction Research: A Systematic Review”. In: *PLOS Medicine* 9.5 (May 2012), pp. 1–13.
- [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] 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.
- [6] 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>.
- [7] Livia Faes et al. “Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?”. In: *Frontiers in Digital Health* 4 (2022).
- [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] Guo Haixiang et al. “Learning from class-imbalanced data: Review of methods and applications”. In: *Expert Systems with Applications* 73 (2017), pp. 220–239.
- [10] 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).
- [11] Bartosz Krawczyk. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.

- [17] 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/>.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Ben Van Calster et al. “Calibration: the Achilles heel of predictive analytics”. In: *BMC Medicine* 17.1 (2019), p. 230.
- [23] 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.
- [24] Laure Wynants et al. “Three myths about risk thresholds for prediction models”. In: *BMC Medicine* 17.1 (2019), p. 192.
- [25] Lian Yu and Nengfeng Zhou. *Survey of Imbalanced Data Methodologies*. 2021. DOI: 10.48550/ARXIV.2104.02240. URL: <https://arxiv.org/abs/2104.02240>.