Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

A demonstration of the LaTeXclass file for Statistics in Medicine with Rmarkdown

Alex Carriero*[1,2] | Maarten van Smeden[2] | Kim Luijken[3] | Ben van Calster[4]

[1]Department of Incredible Research, University A, City A, Country A
[2]Department of Applied Things, University B, City B, Country B
[3]Very Important Stuff Committee, Institute C, City C, Country C
[4]Very Important Stuff Committee, Institute D, City D, Country D

Correspondence
*Corresponding author name, This is sample corresponding address. Email: authorone@gmail.com

Present Address
This is sample for present address text this is sample for present address text

hello its me

KEYWORDS:
Class file; LaTeX; Statist. Med.; Rmarkdown;

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 as a major problem 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].

In a clinical context, a model is only useful if it is well calibrated[@]. This is because in practice, the risk predictions are used by clinicians to council patients and inform treatment decisions. If a model is poorly calibrated, the personal costs to the patient may be enormous. Further, 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.

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. Only one study has assessed the impact of imbalance corrections on model calibration. **(author?)**[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 **(author?)**[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 | METHODS

The performance of several imbalance corrections is compared by means of a simulation study.

2.1 | Imbalance Corrections and Classification Algorithms

Imbalance Corrections

- random under sampling (RUS)[?],
- random over sampling (ROS)[?],
- synthetic majority over sampling technique (SMOTE)[?],
- SMOTE-edited nearest neighbors (SMOTE-ENN)[?].

Classification Algorithms

- logistic regression (base R),

- support vector machine[?],
- random forest[?],
- XG Boost[?],
- RUSBoost[?]
- EasyEnsemble[?]

2.2 | Full Study vs. Pilot Study

Full Study

To gain insight into the performance of imbalance corrections under various scenarios we intend to vary three characteristics of the data generated in our simulation study: number of predictors, event fraction and sample size. In the full study we investigate the effect of imbalance corrections on prediction model performance in 27 (3 x 3 x 3) unique scenarios. 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 presented in **?**. Sample size will then vary through the set $\{\frac{1}{2}$N, N and 2N$\}$.

Pilot Study

In this study we investigate the baseline performance of the prediction models. No imbalance corrections will be applied. The aim in this study is to determine the baseline performance of prediction models developed with the six classification algorithms under 2 scenarios. These scenarios, are presented in Table 1.

2.3 | Simulation Study

We adhere to the ADEMP guidelines for the design and reporting of our simulation study[?].

Aim

We aim to determine the best practices for handling class imbalance when developing clinical prediction models for dichotomous risk prediction. Under a variety of scenarios, four imbalance corrections and six classification algorithms will be used to train prediction models; models will then be systematically compared based on their out-of-sample predictive performance.

We aim to identify any combination of imbalance correction and classification algorithm that, together, produce a model which outperforms it's associated control model (a model trained using the classification algorithm and no imbalance correction).

Data-Generating Mechanism

Data for each class is generated independently from two distinct multivariate normal distributions:

Class 0: $\mathbf{X} \sim mvn(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) = mvn(\mathbf{0}, \boldsymbol{\Sigma_0})$

Class 1: $\mathbf{X} \sim mvn(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) = mvn(\boldsymbol{\Delta_\mu}, \boldsymbol{\Sigma_0} - \boldsymbol{\Delta_\Sigma})$

The parameters (mean vector and covariance matrix) of the data generating distributions are distinct between the classes. In the formulae above, $\boldsymbol{\Delta_\mu}$ refers to the vector housing the difference in predictor means between the two classes. Similarly, $\boldsymbol{\Delta_\Sigma}$ refers to the matrix housing the difference in predictor variances/covariances between the classes.

In class 0, all predictor means are fixed to zero and all variances are fixed to 1. In class 1, all means are stored in the vector $\boldsymbol{\Delta_\mu}$, there is no variation in means among predictors within a class, thus, every element in the vector $\boldsymbol{\Delta_\mu}$ is equivalent; denoted by δ_μ. Similarly, there is no variation in predictor variances within a class, so every diagonal element in $\boldsymbol{\Delta_\Sigma}$ is equivalent; diagonal elements are denoted by δ_Σ.

Finally, 80% of the predictors are allowed to covary. All correlations among predictors in each class are set to 0.2. Correlation matrices between the classes are therefore, equivalent; off-diagonal elements of $\boldsymbol{\Delta_\Sigma}$ are computed such that the correlation matrices between the two classes are equivalent. Note, the covariance matrices are *not* equivalent between the classes. For example, in scenario where we have 8 predictors:

mean and covariance structure for class 0,

$$\boldsymbol{\mu_0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma_0} = \begin{bmatrix} 1 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 \\ 0.2 & 1 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 1 & 0.2 & 0.2 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 1 & 0.2 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

mean and covariance structure for class 1,

$$\boldsymbol{\mu_1} = \begin{bmatrix} \delta_\mu \\ \delta_\mu \\ \delta_\mu \\ \delta_\mu \\ \delta_\mu \\ \delta_\mu \\ \delta_\mu \\ \delta_\mu \end{bmatrix}, \boldsymbol{\Sigma_1} = \begin{bmatrix} 1-\delta_\Sigma & z & z & z & z & z & 0 & 0 \\ z & 1-\delta_\Sigma & z & z & z & z & 0 & 0 \\ z & z & 1-\delta_\Sigma & z & z & z & 0 & 0 \\ z & z & z & 1-\delta_\Sigma & z & z & 0 & 0 \\ z & z & z & z & 1-\delta_\Sigma & z & 0 & 0 \\ z & z & z & z & z & 1-\delta_\Sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-\delta_\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-\delta_\Sigma \end{bmatrix}$$

.

Here, $z = \frac{(1-\delta_\Sigma)*0.2}{1}$, to ensure the correlation matrices of the two classes are equivalent.

For every scenario, Tte parameter values for the data generating distributions (δ_μ and δ_Σ) in each class are selected to generate a ΔC Statistic = 0.85. Their values are computed analytically, based on formulae from **?**. This derivation is shown in Appendix A.

Finally, given that data for each class are generated independently, we have excellent control over how many observations are generated under each class. The number of observations from the positive class (n_1) will be will be sampled from the binomial distribution with probability equal to the specified event fraction. The number of observations in the negative class (n_0) will then be computed as $N - n_1$, where N is the total sample size specified for the prediction model.

Estimands

The focus of this study is the out-of-sample predictive performance of clinical prediction models for dichotomous risk prediction.

Methods

Under each scenario, 2000 data sets will be generated. Each data set will be comprised of training and validation data. The training and validation data will be generated independently using identical data generating mechanisms. This is done to ensure the same event fraction in the training and validation data. The validation data is generated to be 10x larger than the training set.

For each generated data set, five imbalance corrections (four and one control) will be applied to the training set. Six prediction models will then be developed for each of the five imbalance corrected training sets. In other words, each data set will result in: 5 corrected training sets x 6 classification algorithms = 30 prediction models. All models will be trained using training data sets. Out-of-sample performance will be then be assessed using the validation data.

Performance Measures

Out-of-sample model performance will be assessed using measures of discrimination, accuracy and calibration.

Discrimination will be measured by area under the receiver operator curve (ΔC-statistic); computed using the function *auc* from `pROC`[?] .

Accuracy will be measured by Brier Score calculated using the equation (1).

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{1}$$

where, N is the sample size, p_i represents the predicted probability for the i^{th} observation and o_i represents the observed binary outcome (0 or 1) for the i^{th} observation. Measures of accuracy which involve the selection of a decision threshold (e.g., total accuracy, sensitivity, specificity) will not be considered.

Calibration will be measured empirically in terms of calibration intercept and slope. Calibration intercept is calculated as the regression intercept resulting from the regression equation shown in (2). Calibration slope is calculated as the regression slope resulting from the regression equation shown in (3). Model calibration will be visualized using flexible model calibration curves fit using the loess regression.

- calibration intercept and slope formulas
- means and mcmc errors will be presented

Software

All analyses will be conducting using R version 4.1.2[?] . For the full study, the high performance computers at University Medical Center Utrecht will be utilized.

3 | RESULTS

4 | DISCUSSION

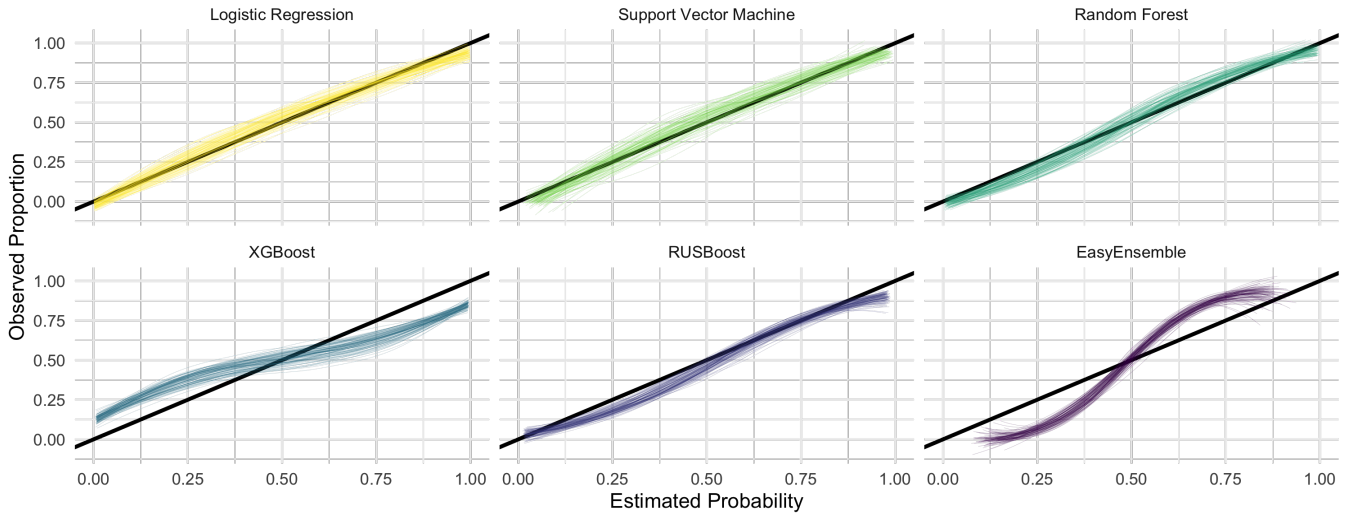
4.1 | Figures and Tables

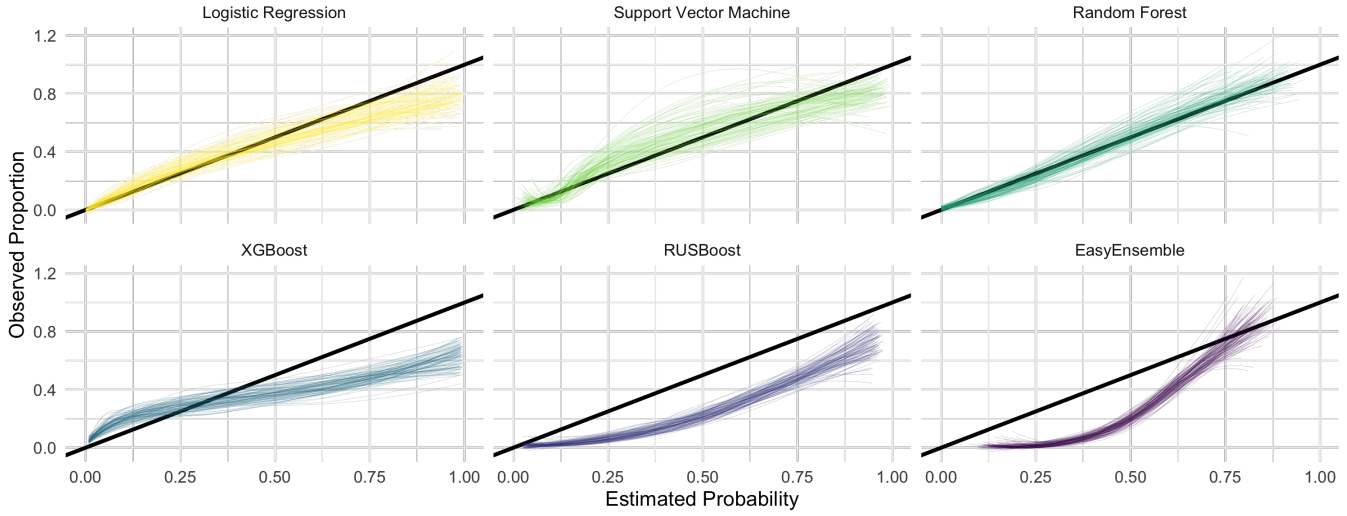
	No. Predictors	Event Fraction	N Level	Sample Size
1	8	0.50	N	385
2	8	0.20	N	247
3	8	0.02	N	1797

TABLE 1 Table Caption

TABLE 2 Nice Table

	Logistic Regression	Support Vector Machine	Random Forest	XGBoost	RUSBoost	EasyEnsemble
Event Fraction: 0.5						
auc_mean	0.84	0.85	0.84	0.79	0.81	0.82
auc_sd	0.01	0.01	0.01	0.01	0.01	0.01
bri_mean	0.16	0.16	0.16	0.20	0.18	0.19
bri_sd	0.00	0.00	0.00	0.01	0.01	0.01
int_mean	-0.01	0.00	-0.02	-0.03	-0.18	0.00
int_sd	0.15	0.15	0.13	0.23	0.10	0.04
slp_mean	0.94	1.04	1.19	0.47	1.06	2.28
slp_sd	0.13	0.16	0.10	0.02	0.08	0.19
Event Fraction: 0.2						
auc_mean	0.84	0.81	0.81	0.76	0.80	0.81
auc_sd	0.01	0.03	0.02	0.02	0.02	0.02
bri_mean	0.12	0.12	0.13	0.15	0.18	0.20
bri_sd	0.01	0.01	0.01	0.01	0.01	0.01
int_mean	0.01	0.02	-0.04	0.05	-1.32	-1.30
int_sd	0.23	0.21	0.19	0.32	0.13	0.08
slp_mean	0.85	1.11	1.08	0.45	1.08	2.31
slp_sd	0.13	0.31	0.12	0.04	0.14	0.27
Event Fraction: 0.02						
auc_mean	0.84	0.70	0.76	0.71	0.78	0.81
auc_sd	0.01	0.03	0.02	0.03	0.03	0.02
bri_mean	0.02	0.02	0.02	0.02	0.16	0.20
bri_sd	0.00	0.00	0.00	0.00	0.02	0.01
int_mean	0.01	-0.01	-0.12	-0.34	-3.72	-3.74
int_sd	0.17	0.17	0.17	0.17	0.17	0.08
slp_mean	0.92	1.52	0.66	0.51	0.85	2.27
slp_sd	0.11	1.05	0.08	0.05	0.17	0.27





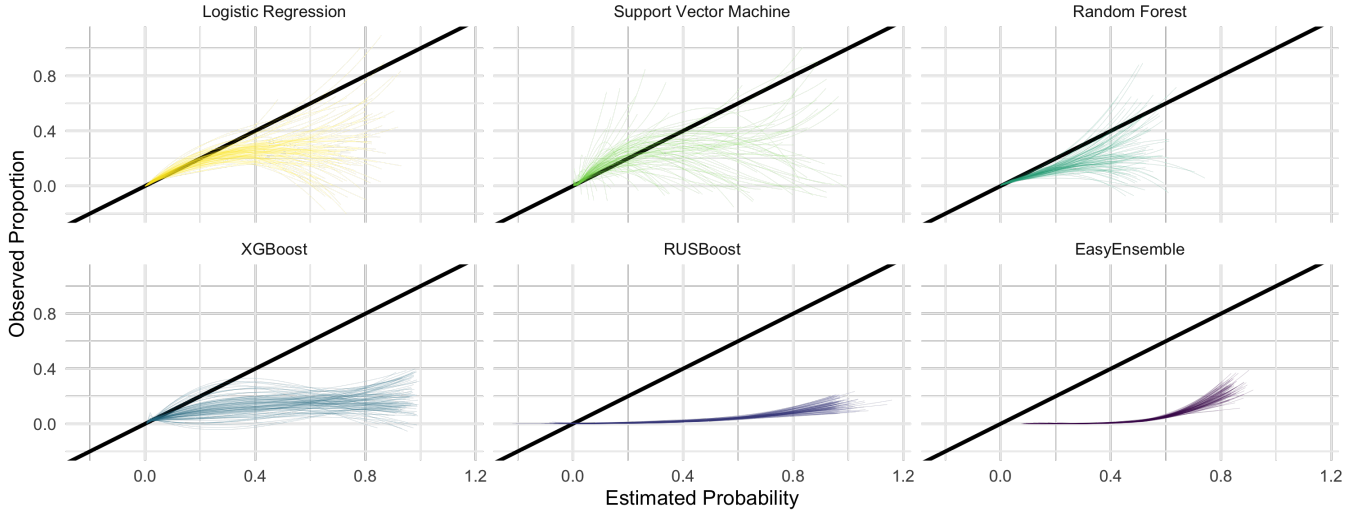


FIGURE 1 Fancy Caption

References

1. Maheshwari S, Jain R, Jadon R. An Insight into Rare Class Problem: Analysis and Potential Solutions. *Journal of Computer Science* 2018; 14(6): 777–792.

2. Yu L, Zhou N. Survey of Imbalanced Data Methodologies. 2021. doi: 10.48550/ARXIV.2104.02240

3. Van Calster B, McLernon DJ, Smeden vM, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine* 2019; 17(1): 230.

4. Branco P, Torgo L, Ribeiro RP. A Survey of Predictive Modeling on Imbalanced Domains. 2016; 49(2). doi: 10.1145/2907070

5. López V, Fernández A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 2013; 250: 113–141.

6. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* 2017; 73: 220-239.

7. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 2016; 5(4): 221–232.

8. Goorbergh v. dR, Smeden vM, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association* 2022; 29(9): 1525–1534. doi: 10.1093/jamia/ocac093