RESEARCH REPORT

Pilot Study: Assessing the impact of class imbalance on the performance of prediction models developed for dichotomous risk prediction.

Alex Carriero
9028757

Supervisors
Dr. Maarten van Smeden
Dr. Kim Luijken
Dr. Ben van Calster

Methodology and Statistics for the Behvavioural, Biomedical and Social Sciences

Utrecht Univeristy
January 2023

Word count: 2488/2500

1. Introduction

Prediction modelling in medicine is receiving increased attention from the clinical community [1]. Most often, the purpose of a clinical prediction model is to estimate a patient's risk of experiencing a particular health event (e.g., disease). Risk estimates are then used to inform clinical decisions; for instance, to decide if a patient is a good candidate for surgery [2,3]. Due to the rare nature of many diseases, data available to train clinical prediction models often exhibit class imbalance (i.e., observations from patients with vs. without the event of interest are not equally represented in the data). When prediction models are trained using imbalanced data, model performance is diminished; performance for the underrepresented class suffers the most [4–6]. Consequently, class imbalance correction methodologies are proposed as a solution [4,7].

While an abundance of imbalance correction methodologies exist [7–9], information regarding the effect of such corrections on model calibration is sparse. Calibration is defined as the accuracy of the risk estimates produced by a prediction model; it measures the agreement between the risk estimates and observed event fractions in the data [3]. Calibration is best evaluated using a calibration plot [10]. If a model is poorly calibrated, it may produce risk estimates that are misleading [3]. For instance, a poorly calibrated model may produce predicted risks that consistently over- or under-estimate true risk or that are too extreme (too close to 0 or 1) or too modest [3]. This can lead to poor treatment decisions or to clinicians communicating false reassurance or hope to patients [3,11]. Therefore, it is vital that model calibration is assessed for prediction models intended for use in clinical settings.

Only one study has assessed the impact of imbalance corrections on model calibration. In this study, the authors demonstrate that class imbalance corrections do more harm than good; implementing imbalance corrections resulted in dramatically deteriorated model calibration, to the point that no corrections were recommended [12]. In this study, prediction models were developed using logistic regression or penalized logistic regression [12]. In practice, most prediction models developed for clinical use do not use regression based methods [13]. Rather, a recent systematic review of clinical prediction models indicated that other classification algorithms, like support vector machine and tree-based methods, are more common [13]. The impact of imbalance corrections on model calibration is currently unknown for prediction models developed using these other classification algorithms.

Motivated by the work of Goorbergh and colleagues [12], 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 wide variety of classification algorithms including: linear classifiers (logistic regression, support vector machine), ensemble learning algorithms (random forest, XGBoost) and algorithms specifically designed to handle class imbalance (RUSBoost, EasyEnsemble). As a first step, we design and implement a pilot study (the focus of this report). In this pilot study, we use a simulation study to illustrate the baseline performance (no imbalance corrections) of prediction models trained in the presence of different class imbalance scenarios. Furthermore, we aim to answer the question: *how does class imbalance affect the performance of clinical prediction models developed for dichotomous risk prediction?*

2. Methods

In this research project, we aim to determine the best practices for handling class imbalance when developing clinical prediction models for dichotomous risk prediction. As a first step, we used a pilot simulation study to illustrate the performance of prediction models in different class imbalance scenarios, without implementing imbalance corrections. In this paper, we present our pilot study; we adhere to the ADEMP guidelines for the design and reporting of our simulation study [14].

Aim

We aimed to assess the out-of-sample predictive performance of prediction models trained with six common classification algorithms, for three class imbalance scenarios. To illustrate the baseline performance of the prediction models for the various class imbalance scenarios, no imbalance corrections were applied to the data before training prediction models.

Data-Generating Mechanism

We generated data for each class independently using two distinct multivariate normal (mvn) distributions:

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

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

Here, class 0 refers to the negative class (non-events) and class 1 refers to the positive class (events). The differences in parameter values between the two classes are represented in the formulae above by $\boldsymbol{\Delta_\mu}$ and $\boldsymbol{\Delta_\Sigma}$; a vector and matrix comprised of the differences in predictor means, and variances/ covariances, between the classes, respectively. We specified no variation in means among predictors within a class, making all elements in the vector $\boldsymbol{\Delta_\mu}$ equivalent; denoted by δ_μ. Similarly, we specified no variation in predictor variances within a class, making all diagonal elements in the matrix $\boldsymbol{\Delta_\Sigma}$ equivalent; diagonal elements are denoted by δ_Σ.

For class 0, we fixed all predictor means to zero and variances to 1. For class 1, all means were non-zero and are represented by the vector $\boldsymbol{\Delta_\mu}$. Finally, we allowed 80% of the predictors to covary. All non-zero correlations among predictors in each class were set to 0.2. To ensure the correlation among predictors was not stronger in one class, we fixed the correlation matrices of the two classes to be equal. This was accomplished by computing the off-diagonal elements of $\boldsymbol{\Delta_\Sigma}$ such that the correlation matrices of the two classes were equivalent. Note, the covariance matrices were *not* equivalent between the classes.

For instance, with 8 predictors:

the mean and covariance structure for class 0 is,

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

and mean and covariance structure for class 1 is,

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

Here, $z = (1-\delta_\Sigma) * 0.2$, to ensure equivalent correlation matrices between the two classes.

Scenarios

We simulated data to reflect three unique class imbalance scenarios (Table 1). This was accomplished by varying the event fraction (proportion of patients with the event of interest) through the set $\{0.5, 0.2, 0.02\}$. For all scenarios, 8 predictors were generated and sample size (N) was determined as the minimum sample size required for the prediction model (given the number of predictors, event fraction and expected concordance statistic, using the R package `pmsampsize`[15].

For every scenario, parameter values for the data generating distributions (δ_μ and δ_Σ) were selected to generate a concordance statistic (C) of 0.85. Under the assumption of normality for all predictors (in each class), the concordance statistic of the data can be expressed as a function of $\boldsymbol{\Delta_\mu}$, $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$ [16]. Optimal values of δ_μ and δ_Σ values were computed analytically, based on the following formula [16]:

$$C = \Phi\left(\sqrt{\boldsymbol{\Delta_\mu}' \left(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1}\right)^{-1} \boldsymbol{\Delta_\mu}}\right). \quad (1)$$

In equation (1), Φ represents the cumulative density function of the standard normal distribution; $\boldsymbol{\Delta_\mu}$, $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$ maintain their previous definitions. To ensure a unique solution, δ_Σ was fixed at 0.3 for each scenario, while equation (1) was solved to yield the appropriate value of δ_μ in each scenario.

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

Parameter values for the data generating distributions of the simulation scenarios are presented in Table 1.

Table 1: Summary of the data generating parameters for each simulation scenario.

Event Fraction	No. Predictors	N	δ_μ	δ_Σ	C
0.50	8	385	0.6043	0.3	0.85
0.20	8	247	0.6043	0.3	0.85
0.02	8	1797	0.6043	0.3	0.85

Simulation Methods

Under each simulation scenario, 200 data sets were generated. Each data set was comprised of training and validation data. The training and validation data were generated independently using identical data generating mechanisms. Validation data sets were generated to be ten times larger than the training data sets.

For each generated data set, six prediction models were developed, each using a different classification algorithm. All prediction models were trained using the training data. Out-of-sample performance was then assessed using the validation data.

Classification algorithms were selected based on a systematic review identifying common algorithms used to develop prediction models in a medical context[13]. These algorithms include: logistic regression, support vector machine, random forest and XGBoost. Additionally, based on literature summarizing common strategies to handle class imbalance[6–8,17], we included two ensemble learning algorithms designed specifically to handle class imbalance: RUSBoost and EasyEnsemble. Classification algorithms were implemented with their default hyper-parameters; no hyper-parameter tuning was conducted. All classification algorithms and the R packages used for their implementation are summarized in Table 2.

Table 2: Summary of classification algorithms used in simulation study.

Classification Algorithm	Abbreviation	R Package
Logistic Regression	LR	base R [18]
Support Vector Machine	SVM	e1071 [19]
Random Forest	RF	randomForest [20]
XGBoost	XG	xgboost [21]
RUSBoost	RB	ebmc [22]
EasyEnsemmble	EE	iric [23]

Performance Measures

Out-of-sample model performance was assessed using measures of calibration, discrimination and overall performance. All performance metrics were computed using validation data.

Calibration was measured using visual and empirical metrics. For each simulation iteration, we fitted a flexible calibration curve for each model; this was done with loess regression using `ggplot2` [24]. In a flexible calibration curve, when estimated probabilities (x-axis) correspond well with the observed proportions in the data (y-axis), the curve follows a diagonal line ($y = x$) [3]. In addition to the calibration plots, calibration intercept and slope were calculated. With respect to calibration intercept and slope, ideal calibration is represented by values of 0 and 1, respectively [25].

Discrimination refers to a model's ability to distinguish between the classes [10]. The concordance statistic was used to measure model discrimination; computed using the R package `pROC` [26]. For dichotomous risk prediction, it is equivalent to the area under the Receiver Operator Characteristic curve [10,25]. A model which perfectly discriminates between the classes will have a concordance statistic of 1; the minimum value for this statistic is 0.5 [25].

Overall performance was measured by the Brier score. This metric reflects both model discrimination and calibration and is calculated according to the following formula [25]:

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

where N is the sample size, p_i represents the estimated probability for the ith individual and o_i represents the observed outcome (0 or 1) for the ith individual. In an ideal model, estimated probabilities approximate the observed outcome well for all individuals; ideal models produce a Brier score near to zero.

For empirical measures of model performance (concordance statistic, Brier score, calibration intercept and calibration slope), the mean over all iterations and corresponding Monte Carlo error were reported.

Software

All analyses were conducting using R version 4.1.2 [18].

3. Results

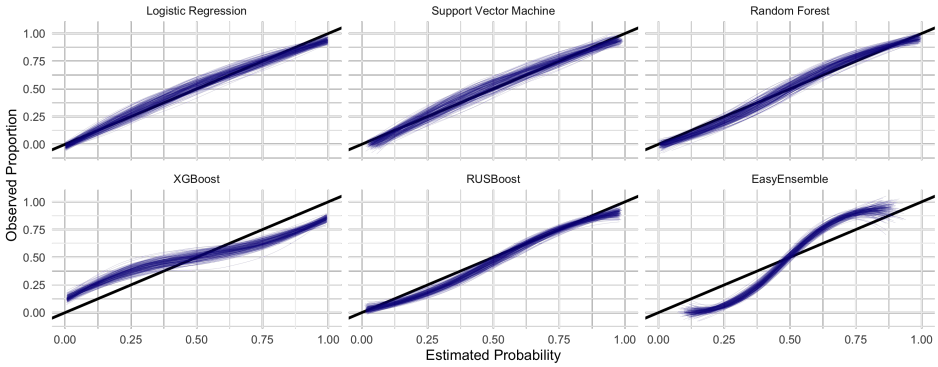
Results are summarized in Table 3 and calibration plots are displayed in Figure 1.

For the class balanced scenario (event fraction = 0.5), all algorithms, except XG and EE, were well calibrated, on average (Figure 1a). While both XG and EE had average calibration intercepts near zero, their average calibration slopes deviated from 1 (Table 3). We see that for XG, the predicted risks above 0.5 overestimated true risk, while the predicted risks below 0.5 underestimated true risk (Figure 1a). In other words, the XG models resulted in risk estimates which were too extreme (calibration slope = 0.464). The opposite pattern was true for EE; EE produced risk estimates which were too moderate (calibration slope = 2.279). In the class balanced scenario, SVM and LR had similar discrimination and overall performance and both outperformed the other algorithms (Table 3).

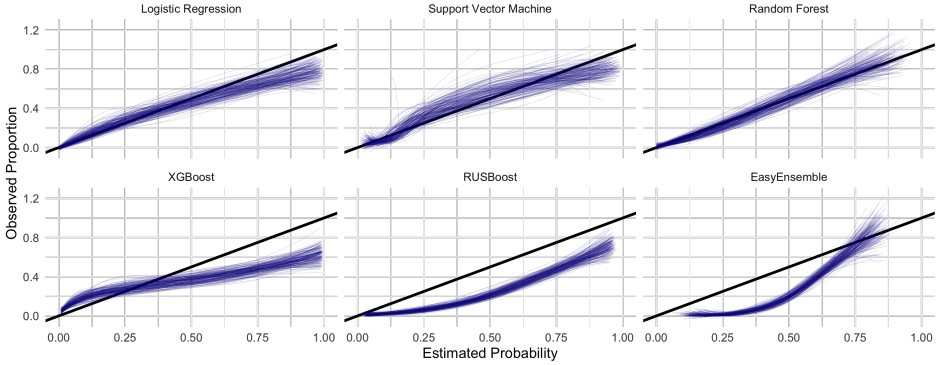
For the moderate class imbalance scenario (event fraction = 0.2), all algorithms exhibited worse calibration, on average, compared to the class balanced scenario. While LR, SVM and RF maintained adequate calibration in this scenario, XG, RB and EE all, on average, produced predicted risks that over-estimated true risk (Figure 1b). In this scenario, we see a slight increase in the variation of the calibration curves produced across the iterations compared to the class balanced scenario (Figure 1b). This was especially apparent for SVM; calibration slope estimates for SVM also varied greatly across the iterations (Table 3). With respect to discrimination and overall performance, in this scenario, LR was the best performing algorithm (Table 3).

In the most extreme class imbalance scenario (event fraction = 0.02), all algorithms exhibited miscalibration. From Figure 1c, we see that for LR, SVM and RF, there was large variation in the calibration curves produced across the simulation iterations. Meanwhile, for XG, RB, and EE, the calibration curves did not vary much across the iterations, rather, they exhibited a specific pattern of miscalibration: all predicted risks over-estimated true risk. With respect to discrimination and overall performance, in this scenario, LR was again, the best performing algorithm (Table 3).

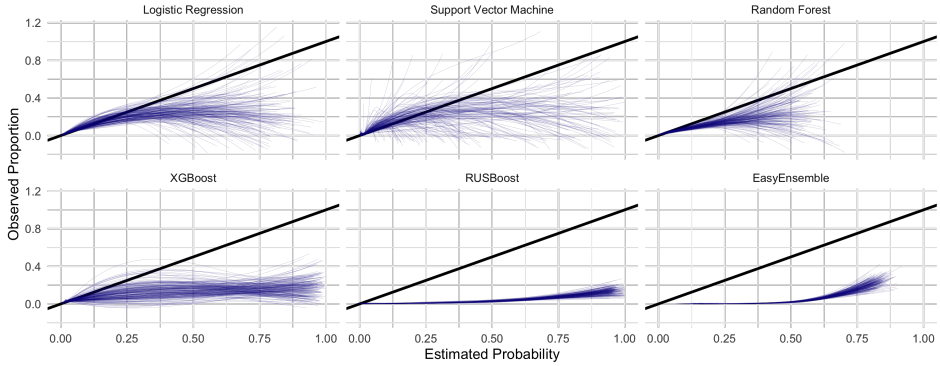
Overall, as imbalance between the classes was magnified, model calibration deteriorated for all algorithms. From Table 3, we also see that discrimination decreased for all algorithms. Interestingly, as imbalance between the classes was magnified, overall performance appeared to improve, especially for models developed with LR, SVM and RF. This apparent improvement in overall performance is misleading and is the result of a poor choice in performance metric.



(a) Flexible calibration curves with event fraction: 0.5



(b) Flexible calibration curves with event fraction: 0.2



(c) Flexible calibration curves with event fraction: 0.02

Figure 1: Visual representation of model calibration for each simulation scenario.

Table 3: Mean (Monte Carlo error) of performance metrics across 200 iterations in each simulation scenario.

	Concordance Statistic	Brier Score	Calibration Intercept	Calibration Slope
Event Fraction: 0.5				
LR	0.845	0.161	-0.001	0.928
	(0.007)	(0.004)	(0.140)	(0.098)
SVM	0.849	0.158	-0.002	1.022
	(0.008)	(0.004)	(0.131)	(0.120)
RF	0.842	0.163	-0.016	1.168
	(0.008)	(0.004)	(0.124)	(0.087)
XG	0.790	0.205	-0.046	0.464
	(0.012)	(0.007)	(0.217)	(0.024)
RB	0.813	0.178	-0.183	1.063
	(0.010)	(0.005)	(0.081)	(0.072)
EE	0.826	0.187	0.001	2.279
	(0.010)	(0.004)	(0.045)	(0.176)
Event Fraction: 0.2				
LR	0.836	0.122	-0.035	0.860
	(0.013)	(0.005)	(0.216)	(0.137)
SVM	0.810	0.123	-0.025	1.049
	(0.026)	(0.007)	(0.197)	(1.057)
RF	0.808	0.125	-0.079	1.079
	(0.018)	(0.005)	(0.179)	(0.120)
XG	0.755	0.153	0.006	0.445
	(0.022)	(0.008)	(0.304)	(0.040)
RB	0.796	0.184	-1.335	1.078
	(0.018)	(0.011)	(0.107)	(0.115)
EE	0.814	0.196	-1.302	2.331
	(0.018)	(0.009)	(0.064)	(0.256)
Event Fraction: 0.02				
LR	0.839	0.019	0.014	0.904
	(0.011)	(0.001)	(0.199)	(0.124)
SVM	0.698	0.019	-0.013	1.255
	(0.034)	(0.001)	(0.193)	(2.652)
RF	0.756	0.019	-0.115	0.655
	(0.022)	(0.001)	(0.189)	(0.083)
XG	0.707	0.022	-0.331	0.508
	(0.028)	(0.001)	(0.186)	(0.047)
RB	0.778	0.158	-3.696	0.869
	(0.027)	(0.019)	(0.157)	(0.152)
EE	0.810	0.200	-3.738	2.295
	(0.019)	(0.014)	(0.080)	(0.304)

4. Discussion

In this paper we investigated the impact of class imbalance on the performance of clinical prediction models developed with six classification algorithms. The results of this study illustrate the performance of these classification algorithms across three class imbalance scenarios; no imbalance corrections were applied to the data before training prediction models. Overall, we saw that as the event fraction was decreased, models exhibited increased miscalibration. At the most extreme event fraction (0.02), models developed with LR, SVM and RF exhibited miscalibration in an unpredictable way. There was large variation among the flexible calibration curves across the simulation iterations; some curves consistently over-estimated true risk while others consistently under-estimated true risk. For models developed with XG, RB, and EE, at the most extreme event fraction, there was a very specific pattern of miscalibration; all over-estimated true risk. Overall, we demonstrated that as class imbalance increased, both calibration and discrimination decreased, for all prediction models considered.

We note two significant limitations to this study. First, Brier score appeared to be an uninformative measure of overall performance when class imbalance was extreme. With an event fraction of 0.02, a trivial majority classifier (a model that predicts everyone will belong to the larger class) would yield a Brier score of 0.02. Therefore, in our future work, we will utilize another metric of overall performance, such a re-scaled Brier score, which is known to be more informative in the presence of class imbalance [25]. Second, models developed with classification algorithms other than logistic regression may have performed worse than expected due to the lack of hyper-parameter tuning. In particular, RB and EE preformed substantially worse than expected. These algorithms are designed to handle class imbalance, yet, they had worse overall performance than a trivial majority classifier at the most extreme event fraction. The relatively poor performance of these algorithms may be due to the lack of hyper-parameter tuning, therefore, future work will allow for hyper-parameter tuning.

Class imbalance is common in medical data sets and in this pilot study we have demonstrated that prediction models may be miscalibrated in the presence of extreme class imbalance. Future work will investigate the best practices for handling class imbalance without compromising model calibration. Goorbergh and colleagues [12] have demonstrated that imbalance corrections may do more harm than good with respect to model calibration for prediction models developed using logistic regression [12]. In our future work we will extend this research by assessing the impact of imbalance corrections and re-calibration procedures on prediction models developed using the wide variety of algorithms considered in this pilot study.

Session Info

```
sessionInfo()
```

```
R version 4.1.2 (2021-11-01)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Monterey 12.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] knitr_1.40       kableExtra_1.3.4 forcats_0.5.1    stringr_1.4.0
 [5] dplyr_1.0.10     purrr_0.3.4      readr_2.1.2      tidyr_1.2.1
 [9] tibble_3.1.6     ggplot2_3.4.0    tidyverse_1.3.2

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.9          svglite_2.1.0       lubridate_1.8.0
 [4] png_0.1-8           assertthat_0.2.1    digest_0.6.29
 [7] utf8_1.2.2          R6_2.5.1            cellranger_1.1.0
[10] backports_1.4.1     reprex_2.0.1        evaluate_0.18
[13] httr_1.4.2          pillar_1.7.0        rlang_1.0.6
[16] googlesheets4_1.0.0 readxl_1.3.1        rstudioapi_0.14
[19] rmarkdown_2.18      webshot_0.5.2       googledrive_2.0.0
[22] munsell_0.5.0       broom_0.7.12        compiler_4.1.2
[25] modelr_0.1.8        xfun_0.34           pkgconfig_2.0.3
[28] systemfonts_1.0.4   htmltools_0.5.4     tidyselect_1.2.0
[31] fansi_1.0.2         viridisLite_0.4.0   crayon_1.5.0
[34] tzdb_0.2.0          dbplyr_2.1.1        withr_2.5.0
[37] grid_4.1.2          jsonlite_1.8.4      gtable_0.3.0
[40] lifecycle_1.0.3     DBI_1.1.2           magrittr_2.0.3
[43] scales_1.2.1        cli_3.4.1           stringi_1.7.6
[46] fs_1.5.2            xml2_1.3.3          ellipsis_0.3.2
[49] generics_0.1.3      vctrs_0.5.1         tools_4.1.2
[52] glue_1.6.2          hms_1.1.1           fastmap_1.1.0
[55] yaml_2.3.6          colorspace_2.0-2    gargle_1.2.0
[58] rvest_1.0.2         haven_2.4.3
```

References

[1] Sandra Eloranta and Magnus Boman. Predictive models for clinical decision making: Deep dives in practical machine learning. *Journal of Internal Medicine*, 292(2):278–295, 2022.

[2] Lingxiao Chen. Overview of clinical prediction models. *Annals of Translational Medicine*, 8(4), 2019.

[3] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra Macaskill, David J. McLernon, Karel G. M. Moons, Ewout W. Steyerberg, Andrew J. Vickers, On behalf of Topic Group 'Evaluating diagnostic tests, and prediction models'of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

[4] Fadel M. Megahed, Ying-Ju Chen, Aly Megahed, Yuya Ong, Naomi Altman, and Martin Krzywinski. The class imbalance problem. *Nature Methods*, 18(11):1270–1272, 2021.

[5] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[6] Lian Yu and Nengfeng Zhou. Survey of imbalanced data methodologies, 2021.

[7] Satyam Maheshwari, R.C. Jain, and R.S. Jadon. An insight into rare class problem: Analysis and potential solutions. *Journal of Computer Science*, 14(6):777–792, May 2018.

[8] Victoria López, Alberto Fernández, Jose G. Moreno-Torres, and Francisco Herrera. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608, 2012.

[9] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[10] Anne A. H. de Hond, Artuur M. Leeuwenberg, Lotty Hooft, Ilse M. J. Kant, Steven W. J. Nijman, Hendrikus J. A. van Os, Jiska J. Aardoom, Thomas P. A. Debray, Ewoud Schuit, Maarten van Smeden, Johannes B. Reitsma, Ewout W. Steyerberg, Niels H. Chavannes, and Karel G. M. Moons. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine*, 5(1):2, 2022.

[11] Richard D Riley, Tim J Cole, Jon Deeks, Jamie J Kirkham, Julie Morris, Rafael Perera, Angie Wade, and Gary S Collins. On the 12th day of christmas, a statistician sent to me . . . *BMJ*, 379, 2022.

[12] Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 06 2022.

[13] Constanza Navarro, Johanna Damen, Maarten van Smeden, Toshihiko Takada, Steven Nijman, Paula Dhiman, Jie Ma, Gary Collins, Ram Bajpai, Richard Riley, Karel Moons, and Lotty Hooft. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, 11 2022.

[14] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

[15] Joie Ensor, Emma C. Martin, and Richard D. Riley. *pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model*, 2022. R package version 1.1.2.

[16] Olga V. Demler, Michael J. Pencina, and Ralph B. D'Agostino Sr. Equivalence of improvement in area under roc curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine*, 30(12):1410–1418, 2011.

[17] Prabhjot Kaur and Anjana Gosain. Empirical assessment of ensemble based approaches to classify imbalanced data in binary classification. *International Journal of Advanced Computer Science and Applications*, 2019.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[19] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2022. R package version 1.7-12.

[20] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[21] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, and Jiaming Yuan. *xgboost: Extreme Gradient Boosting*, 2022. R package version 1.6.0.1.

[22] Hsiang Hao and Chen. *ebmc: Ensemble-Based Methods for Class Imbalance Problem*, 2022. R package version 1.0.1.

[23] Bing Zhu, Zihan Gao, Junkai Zhao, and Seppe K.L.M. vanden Broucke. Iric: An r library for binary imbalanced classification. *SoftwareX*, 10:100341, 2019.

[24] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[25] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, 01 2010.

[26] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.