Simulation Study Protocol

Simulation(s) to assess the impact of class imbalance corrections on the calibration of clinical prediction models.

Alex Carriero

November 27, 2022

1 ADEMP

1.1 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 realistic 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, create a model which outperforms the associated control model (a model trained using the classification algorithm and no imbalance correction). In particular, we aim to determine if any imbalance corrections lead to improved model performance without compromising model calibration.

1.2 Data-Generating Mechanisms

1.2.1 Scenarios

Imbalanced data will be simulated to reflect 27 (3 x 3 x 3) unique scenarios. This is achieved by varying the following three characteristics of the data: 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 presented in Riley et al. (2020). Sample size will then vary through the set {$\frac{1}{2}$N, N and 2N}.

Table 1: Summary of factors to be varied in data simulation.

Factor	**Levels**
No. of predictors	8, 16, 32
Event fraction	0.5, 0.2, 0.02
Sample Size	$\frac{1}{2}N$, N, $2N$

* N represents the minimum sample size for the prediction model.

Under each scenario, 2000 data sets will be generated. Data sets will be comprised of training and test data such that the training data set is 10x larger than the test set.

1.2.2 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 two 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 each scenario, the 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 equation (2) shown in Appendix A. For each simulation scenario, the parameters of the data generating distribution are shown in the Table 2.

Table 2: Summary of parameters used in data generating mechanism for all simulation scenarios

No. Predictors	Event Fraction	Sample Size	δ_μ	δ_Σ	AUC
0.5N					
8	0.50	193	0.60	0.3	0.85
8	0.20	124	0.60	0.3	0.85
8	0.02	899	0.60	0.3	0.85
16	0.50	193	0.49	0.3	0.85
16	0.20	247	0.49	0.3	0.85
16	0.02	1797	0.49	0.3	0.85
28	0.50	296	0.42	0.3	0.85
28	0.20	432	0.42	0.3	0.85
28	0.02	3144	0.42	0.3	0.85
N					
8	0.50	385	0.60	0.3	0.85
8	0.20	247	0.60	0.3	0.85
8	0.02	1797	0.60	0.3	0.85
16	0.50	385	0.49	0.3	0.85
16	0.20	493	0.49	0.3	0.85
16	0.02	3593	0.49	0.3	0.85
28	0.50	592	0.42	0.3	0.85
28	0.20	863	0.42	0.3	0.85
28	0.02	6288	0.42	0.3	0.85
2N					
8	0.50	770	0.60	0.3	0.85
8	0.20	494	0.60	0.3	0.85
8	0.02	3594	0.60	0.3	0.85
16	0.50	770	0.49	0.3	0.85
16	0.20	986	0.49	0.3	0.85
16	0.02	7186	0.49	0.3	0.85
28	0.50	1184	0.42	0.3	0.85
28	0.20	1726	0.42	0.3	0.85
28	0.02	12576	0.42	0.3	0.85

1.2.3 Outcomes

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 event fraction. The number of observations in the negative class (n_0) will then be computed as $N - n_1$.

1.3 Estimands

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

1.4 Methods

To investigate the effect of common class imbalance corrections on model performance, a full-factorial simulation design will be implemented. Four imbalance corrections and one control (no correction) will be implemented for each of six classification algorithms. The classification algorithms and imbalance corrections we will include in our simulation are detailed in Tables 3 and 4 respectively.

Table 3: Summary of class imbalance corrections to be implemented.

Imbalance Correction	**R Package**	**Python Library**
Random Under Sampling	ROSE	imblearn
Random Over Sampling	ROSE	imblearn
SMOTE	smotefamily	imblearn
SMOTE-ENN	*IRIC	imblearn
None	—	—

[*] IRIC package not available on CRAN

Table 4: Summary of classification algorithms to be implemented.

Method	**R Package**	**Python Library**
Logistic Regression	glmnet	scikit-learn
Support Vector Machine	e1701	scikit-learn
Random Forest	randomForest	scikit-learn
XG Boost	xgboost	xgboost
RUSBoost	ebmc	imblearn
EasyEnsemble	*IRIC	imblearn

[*] IRIC package not available on CRAN

In summary, 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 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 test data.

1.5 Performance Measures

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

1.5.1 Discrimination

Discrimination will be measured by area under the receiver operator curve (ΔC-statistic).

1.5.2 Accuracy

Accuracy will be measured by Brier Score:

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

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. Brier score is equivalent to the mean square error between the predicted probabilities and observed outcomes.

Measures of accuracy which involve the selection of a decision threshold (e.g., total accuracy, sensitivity, specificity) will not be considered.

1.5.3 Calibration

Calibration will be measured in terms of calibration intercept and slope. Model calibration will be visualized using flexible model calibration curves fit using the loess function. A sample of 200 calibration curves will be plotted per scenario.

2 Error Handling

It is possible that some models do not converge, especially in the rare event that too few cases are generated in the positive class. Therefore, a catch function will be used during the simulation such that any warnings and messages the models are saved.

Appendix A

Under the assumption of normality for all predictors (in each class), AUC can be calculated directly, using equation (1) (Demler, Pencina, and D'Agostino Sr. 2011). This equation is suitable when the covariance matrices of each class are *not* equivalent. For p predictors, $\boldsymbol{\Delta_\mu}$ is a p x 1 vector housing the difference in predictor means between class 0 and class 1. $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$ represent the covariance matrices of class 0 and 1 respectively and Φ represents the cumulative density function (cdf) of the standard normal distribution.

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

In our project, the differences in predictor means between the classes are equivalent. In other words, the elements of $\boldsymbol{\Delta_\mu}$ are equivalent; denoted by δ_μ. The differences in predictor variances between the classes are also equivalent. The diagonal elements of $\boldsymbol{\Sigma_1}$ are all equal to $(1-\delta_\Sigma)$ as shown in section 1.2.2. To ensure this equation has a unique solution, δ_Σ is fixed to 0.3. Then, we may solve equation (1) to determine the value of δ_μ which yields the desired AUC.

Let $\boldsymbol{A} = (\boldsymbol{\Sigma_0}+\boldsymbol{\Sigma_1})^{-1}$,

$$(\Phi^{-1}(AUC))^2 = \boldsymbol{\Delta_\mu}'\boldsymbol{A}\ \boldsymbol{\Delta_\mu}$$

$$(\Phi^{-1}(AUC))^2 = \begin{bmatrix}\delta_\mu & \delta_\mu & \dots & \delta_\mu & \delta_\mu\end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ \vdots & & \ddots & \\ a_{p1} & a_{p2} & \dots & a_{pp}\end{bmatrix} \begin{bmatrix}\delta_\mu \\ \delta_\mu \\ \vdots \\ \delta_\mu \\ \delta_\mu\end{bmatrix}$$

$$(\Phi^{-1}(AUC))^2 = \delta_\mu^2 \sum_{j=1}^{p}\sum_{i=1}^{p} a_{ij}$$

Based on a desired AUC of 0.85,

$$\delta_\mu = \frac{\Phi^{-1}(0.85)}{\sqrt{\sum_{j=1}^{p}\sum_{i=1}^{p} a_{ij}}}. \tag{2}$$

Equation (2) will be used to derive the appropriate δ_μ for each simulation scenario. Meanwhile, δ_Σ will remain fixed at 0.3 for each scenario.

References

Demler, Olga V., Michael J. Pencina, and Ralph B. D'Agostino Sr. 2011. "Equivalence of Improvement in Area Under ROC Curve and Linear Discriminant Analysis Coefficient Under Assumption of Normality." *Statistics in Medicine* 30 (12): 1410–18.

Riley, Richard D, Joie Ensor, Kym I E Snell, Frank E Harrell, Glen P Martin, Johannes B Reitsma, Karel G M Moons, Gary Collins, and Maarten van Smeden. 2020. "Calculating the Sample Size Required for Developing a Clinical Prediction Model." *BMJ* 368.