Simulation Study Protocol

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

Alex Carriero

November 17, 2022

1 ADEMP

1.1 Aim

We aim to determine the best practices for handling class imbalance when developing clinical prediction models developed 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 can outperform a control model (model developed using the given classification algorithm with no imbalance correction). In particular, we aim to determine which (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 properties 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.

1.2.1 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, $\boldsymbol{\Delta}_{\boldsymbol{\mu}}$ refers to the vector housing the difference in predictor means between the two classes. Similarly, $\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}$ refers to the matrix housing the difference in predictors variances/covariances between the two classes. $\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}$ is a diagonal matrix; all predictor covariances will be equal between the classes.

The parameter values for the data generating distributions in each class are selected to generated a ΔC Statistic = **INSERT**. For each scenario, the parameters of the data generating distribution are included in the Table 2. Mean and standard deviation estimates of AUC are calculated based on a small simulation, in which 2000 data sets are generated. This is done to detail the expected mean and variation of AUC of data generated for each scenario in the full simulation study.

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

Event	No.	Delta Mean	Delta Var	AUC	SD
Fraction	Predictors				
0.5N					
0.50	8	0	0	0	0
0.20	8	0	0	0	0
0.02	8	0	0	0	0
0.50	16	0	0	0	0
0.20	16	0	0	0	0
0.02	16	0	0	0	0
0.50	32	0	0	0	0
0.20	32	0	0	0	0
0.02	32	0	0	0	0
N					
0.50	8	0	0	0	0
0.20	8	0	0	0	0
0.02	8	0	0	0	0
0.50	16	0	0	0	0
0.20	16	0	0	0	0
0.02	16	0	0	0	0
0.50	32	0	0	0	0
0.20	32	0	0	0	0
0.02	32	0	0	0	0
2N					
0.50	8	0	0	0	0
0.20	8	0	0	0	0
0.02	8	0	0	0	0
0.50	16	0	0	0	0
0.20	16	0	0	0	0
0.02	16	0	0	0	0
0.50	32	0	0	0	0
0.20	32	0	0	0	0
0.02	32	0	0	0	0

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 data set, five imbalance corrections (four + one control) will be applied to the training set. Subsequently, six prediction models will 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.

Discrimination:

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

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). 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.

Calibration:

Calibration will be measured in terms of calibration intercept and slope. Model calibration will be visualized using flexible model calibration curves.

2 Error Handling

References

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.