MSR120

Estimating Sample Size For Training Ensemble Machine Learning Models

Nicholas Mitsakakis, PhD, Dan Liu, PhD, Khaled El Emam, PhD
Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada.

BACKGROUND

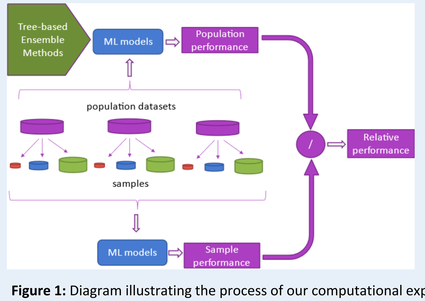
- Machine Learning (ML) models trained on small datasets lead to model instability, overfitting, and poor generalization, resulting in wasted resources and ethical concerns.
- Determining the sample size required for adequate training of ML models is crucial for achieving study goals.
- Current practises include “convenience” sample sizes, oversimplified "rules of thumb" or using methods developed for statistical regression models (e.g. Riley et al. (2020)), which may not be suitable for ML models.
- There is currently a large knowledge gap and lack of guidance in determining sample size requirements for studies using ML methods.

STUDY OBJECTIVE

- To develop an empirically derived sample size calculator for ensemble ML binary classification methods: random forest (RF), Light Gradient Boosting Machine (LGBM) and Extreme Gradient Boosting (XGBoost).

METHODS – SIMULATION DESIGN

- Models were trained, tuned and assessed using 13 large real datasets as population data, obtaining the optimal *population* level performance.
- Population sets were split into 70-30% for training and testing, performance was assessed using the Area Under the ROC Curve (AUC).
- Samples were drawn from each of the training partitions, with sizes based on a geometric series: $b \sim N(\mu = 1.5, \sigma^2 = 0.005^2)$, $n_i = \left[b^{i+9}\right], i = 1, \dots, 23$, where [.] denotes rounding to the closest integer.
- One hundred (100) series were generated for each dataset.
- ML models were trained, tuned and assessed using the samples.
- Tuning used 5-fold cross-validation and Bayesian Optimization.



Sample-trained model performance was compared with population performance to construct the *certainty curve*

Figure 1: Diagram illustrating the process of our computational experiments.

METHODS – CERTAINTY CURVE

- *Certainty curve* C(n) gives the probability for the AUC performance of an ML model trained on data S of size n to exceed a threshold λ relative to population performance AUC(P), i.e.

$$C(n) = Pr_{|S|=n, S \subset P}\left(AUC(S) \geq \lambda \cdot AUC(P)\right)$$

- Given C(n), the required sample size n* for achieving certainty C* (e.g. 80%) is given by n* = min{n: C(n) ≥ C*}.
- Certainty curve is estimated from data from our experiments, using

$$y_{ijk} = I\left(AUC\left(S_{ijk}\right) \geq \lambda \cdot AUC(P_k)\right), i = 1, \dots 23, j = 1, \dots, 100, k = 1, \dots, 13$$

- For each combination of modeling method (LGBM, RF, XGBoost) and value of λ (0.8, 0.85, 0.9) an LGBM model was fitted using *y* as response and *n*, *imbalance factor*, *average standardized entropy* and *total degrees of freedom* as predictors, over all datasets and samples.
- Standardized entropy of a categorical variable with z levels is given by $-\frac{\sum_{i=1}^{z} p_i \log_2(p_i)}{\log_2(z)}$, where p_i is the proportion for level *i.*

PERFORMANCE EVALUATION

- Certainty curve method was evaluated with the percentage median relative error (*mRE*) between the observed and predicted required sample sizes for all 13 datasets:

$$mRE = 100 \cdot median\left(\frac{\hat{n}_i - n_i^*}{n_i^*}\right)$$

- Predicted sample sizes $\hat{n}$ were obtained using leave-one-dataset out, for 80% and 90% certainty.
- Observed sample sizes n^* were obtained with the use of the “observed certainty curve”, estimated by the logistic model logit(C(*n*)) = *a*+*b*·log(*n*), for each dataset separately.

RESULTS

Model	Lambda	Certainty curve (%)	Riley (%)	EPV300 (%)	EPV15 (%)	EPV10 (%)
LGBM	0.80	-66.8	9,012	142,094.5	7,009.9	4,639.9
LGBM	0.85	16.5	4,270.3	142,094.5	7,009.9	4,639.9
LGBM	0.90	56.2	2,868.2	142,094.5	7,009.9	4,639.9
RF	0.80	-37.4	93,404	732,748.5	36,543.1	24,328.5
RF	0.85	8.8	57,824	732,748.5	36,543.1	24,328.5
RF	0.90	130.8	38,930	732,748.5	36,543.1	24,328.5
XGB	0.80	-33	19,078.9	134,669.5	6,638.9	4,392.6
XGB	0.85	56.7	11,970.5	134,669.5	6,638.9	4,392.6
XGB	0.90	395	8,108.9	134,669.5	6,638.9	4,392.6

Model	Lambda	Certainty curve (%)	Riley (%)	EPV300 (%)	EPV15 (%)	EPV10 (%)
LGBM	0.80	-61.3	9,012	142,094.5	7,009.9	4,639.9
LGBM	0.85	25.3	4,270.3	142,094.5	7,009.9	4,639.9
LGBM	0.90	64.8	2,868.2	142,094.5	7,009.9	4,639.9
RF	0.80	-28.1	91,072.0	680,402.1	33,925.7	22,583.6
RF	0.85	9.3	55,466.0	680,402.1	33,925.7	22,583.6
RF	0.90	137.4	36,958.0	680,402.1	33,925.7	22,583.6
XGB	0.80	-22.7	19,078.9	134,669.5	6,638.9	4,392.6
XGB	0.85	74.6	11,970.5	134,669.5	6,638.9	4,392.6
XGB	0.90	471.2	8,108.9	134,669.5	6,638.9	4,392.6

Tables 1,2: Comparison of the percentage median relative error (*mRE*) using leave-one-dataset-out at 80% (left) and 90% (right) certainty, between the certainty curve approach and other methods (Riley et al, 300 Events Per Variable, 15 Events Per Variable, 10 Events Per Variable)

RESULTS

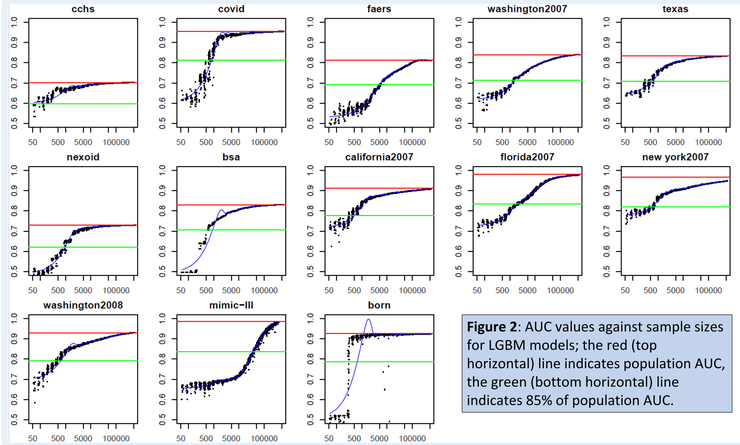


Figure 2: AUC values against sample sizes for LGBM models; the red (top horizontal) line indicates population AUC, the green (bottom horizontal) line indicates 85% of population AUC.

SENSITIVITY ANALYSIS

Discrepancy	Deviation in Mean Standardized Entropy					
	-0.5	-0.25	-0.1	0.1	0.25	0.5
median	48	2	0	-1	-1	-1
within 10%	17.6	47	64.7	41.1	41.1	29.4
within 25%	23.5	52.9	82.3	70.5	76.4	70.5
within 50%	58.8	82.35	88.2	100	100	100

Table 3: Aggregate discrepancy (as percentage relative difference from originally predicted value) for calculated sample size, for different levels of deviation of mean standardized entropy.

- We investigated the impact of accurate standardized mean entropy estimates on sample size calculations using 17 medium-sized datasets with binary outcomes.
- Estimates for mean standardized entropy, degrees of freedom, and imbalance factor were calculated from the data and used to determine sample size with lambda = 0.85 and certainty of 0.8.
- Sample size predictions were recalculated using deviated entropy values to measure the percent relative error.
- An aggregate measure of discrepancy was calculated as the proportion of datasets where the relative error did not exceed 10%, 25%, and 50%.

CONCLUSION AND NEXT STEPS

- We developed a sample size calculator using empirical evidence from a comprehensive simulation study, for three popular tree-based ensemble ML methods (LGBM, RF, XGBoost).
- Our calculator clearly outperforms other methods often used for estimating sample size for ML studies.
- Our methodology can guide prioritization for HEOR studies that use ML models, ensuring efficient resource allocation for informing policy decision making.

REFERENCES AND CONTACT

- Use the QR code for references and supplementary material
- Questions? Email: n.mitsakakis@cheo.on.ca