A study on "The Harm of class imbalance corrections for risk prediction models"

From Journal of the American Medical Informatics Association, 29(9), 2022, 1525–1534 https://doi.org/10.1093/jamia/ocac093 Advance Access Publication Date: 10 June 2022 Research and Applications

Velardita Michele

Statistics for Data Science A/A 22-23

Department of Computer Science, University of Pisa

Project workflow

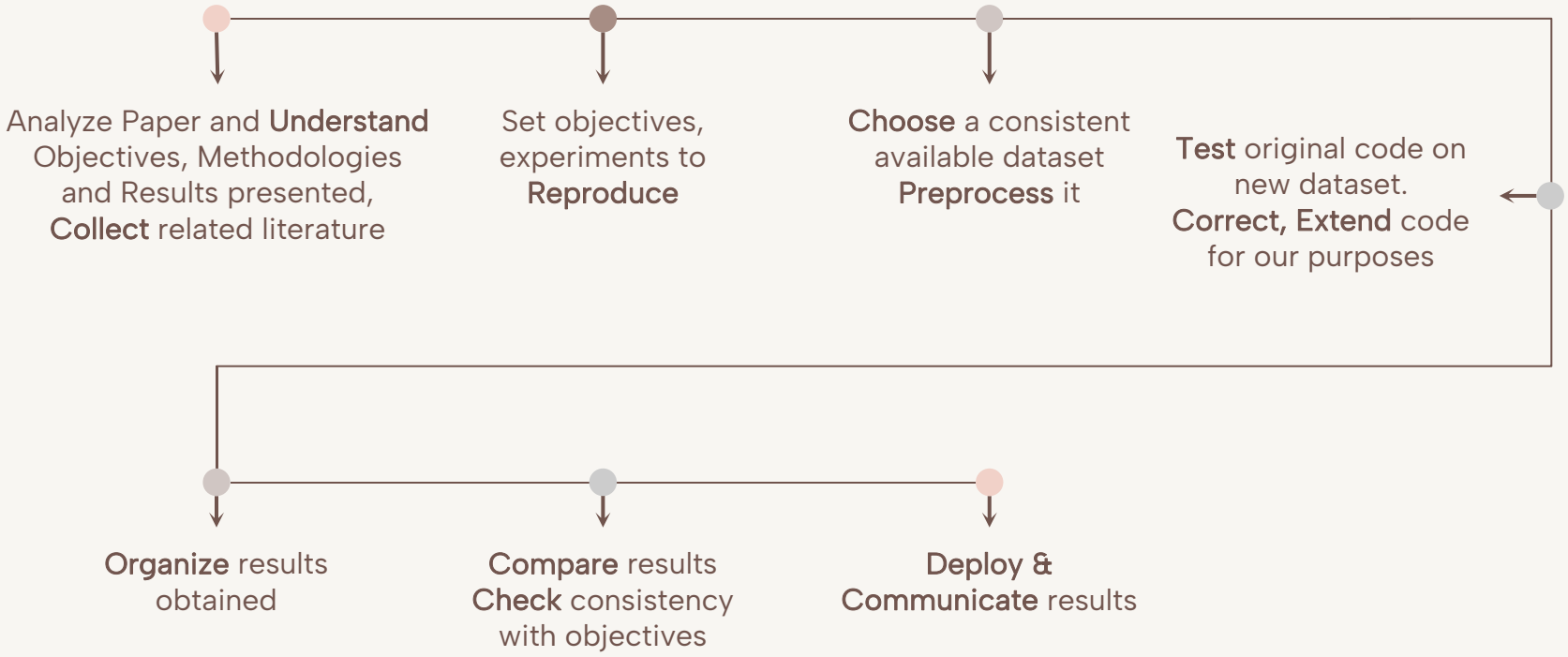


Table of contents

01 - Topic presentation

Problem

Rebalance unbalanced dataset may worsen predictive model performances.

Minority class overestimation

Risk of overtreatment.

Hypothesis

(1) imbalance corrections distort models calibrations

(2) shifting probability threshold has similar impact on sensitivity and specificity as using imbalance correction methods

Case study

Estimate ovarian cancer malignancy probability using an imbalanced dataset.

Unbalance 20%.

Analisys of the performance and clinical utility.

02 - Original Experiments & Results

Dataset: from International Ovarian Tumor Analysis 1999 - 2012

3369 records, not available for privacy issues.

Rebalance techniques

Uncorrected

Random Undersampling (RUS)

Random Oversampling (ROS)

SMOTE

Models

Standard Logistic Regression (SLR)

Penalized Ridge Logistic Regression (L2)

Selected Predictors

Age

Maximum diameter of lesion

Number of papillary structures

Performance Measures

Discrimination: AUROC

Calibration: Reliability of predictions

Classification: Accuracy, Sensitivity, Specificity

Clinical Utility: Net Benefit

02 - Original Experiments & Results

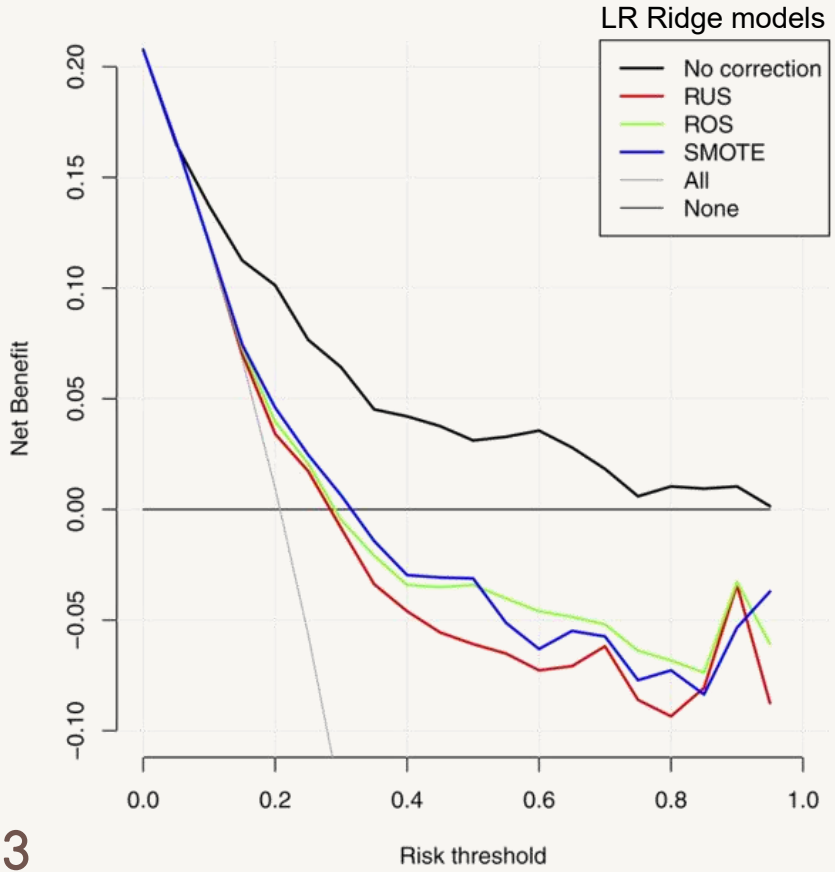
- **Clinical utility** of model in treatment decisions, while taking misclassification errors into account

- **Net Benefit** = $\frac{TP - FP\frac{t}{(1-t)}}{N}$

  Links t and misclassification errors

- **Risk Threshold t :** to select individuals for treatment

- **Default strategies:** treating none
  treating everyone



03 – Our Dataset & Experiments

Dataset: Framingham_heart disease 1948 - 2000s

Cardiovascular study on residents of Framingham, Massachusetts.

4,240 records , 15 attributes, public availability, approved by reliable bodies

Rebalance techniques

Uncorrected

RUS

ROS

SMOTE

ADASYN

Models

SLR

L2

Selected Predictors (3,6,8)

Age

Sys BP

Dia BP

Glucose

TotChol

BMI

CigsPerDay

HeartRate

Performance Measures

Calibration

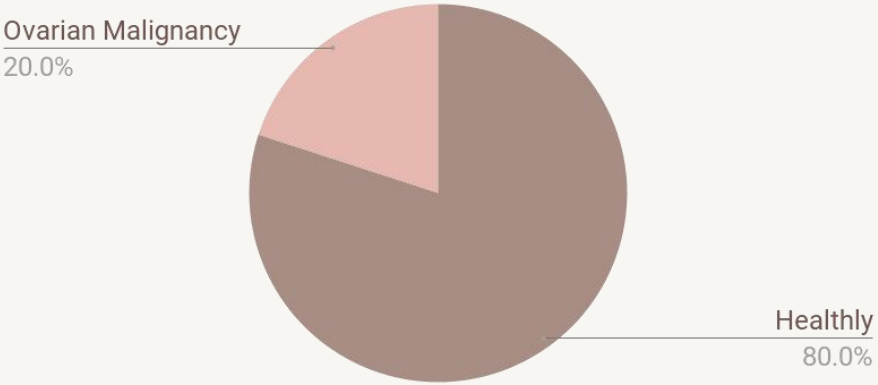
Classification

Clinical Utility

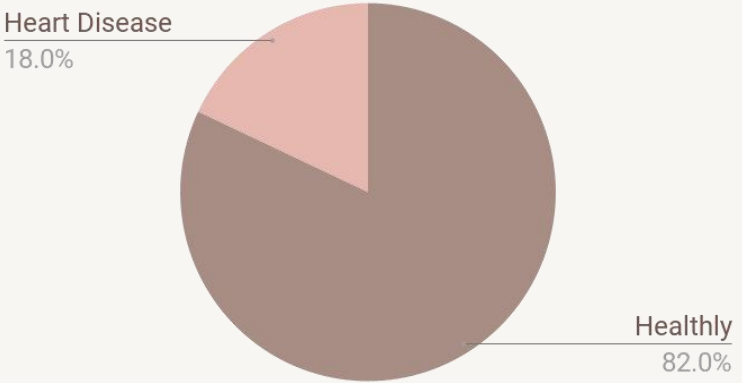
03 – Imbalancing

Imbalancing

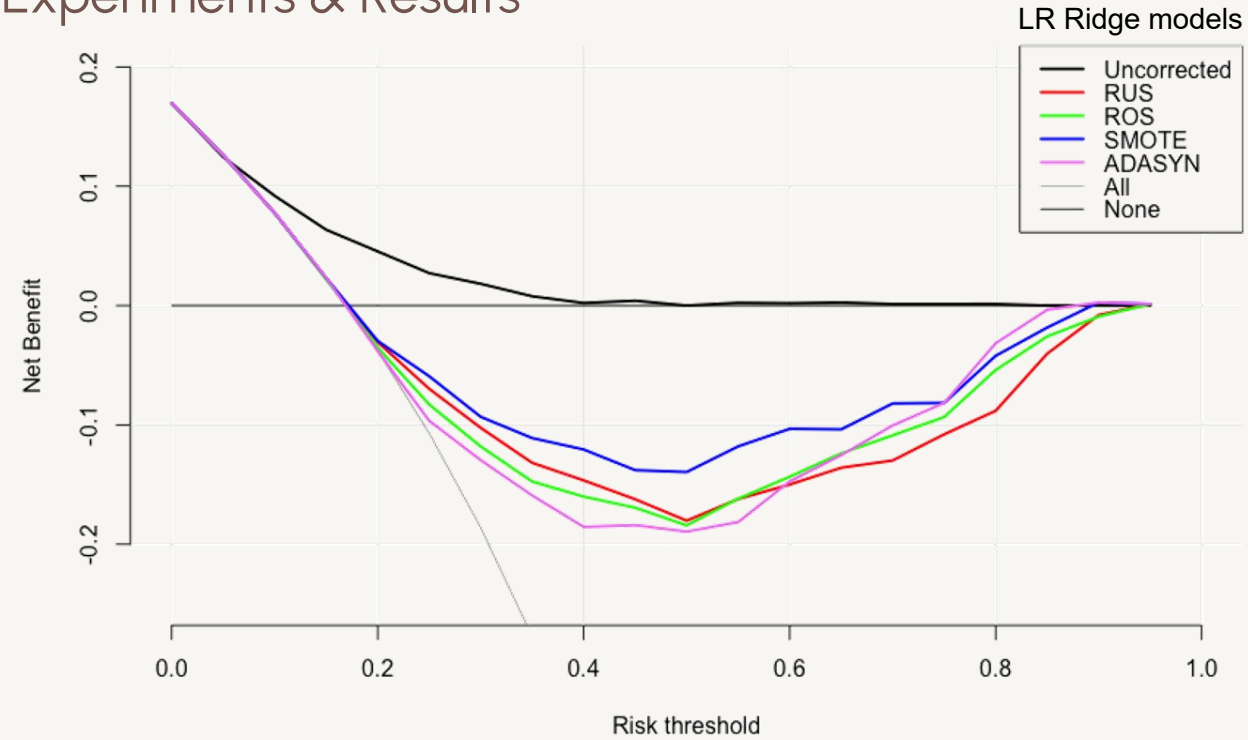
Ovarian cancer dataset



Coronary heart desease dataset



03 - Our Experiments & Results

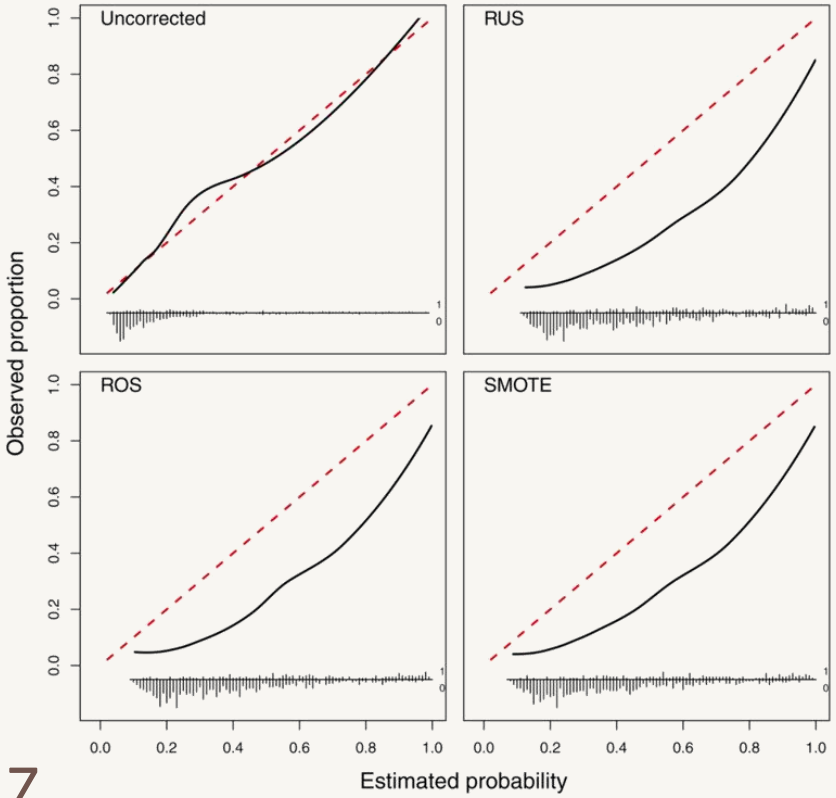


- Consistent with original results
- ADASYN doesn't improve Net Benefit

04 – Comparison

In ovarian cancer dataset, imbalance correction methods yield to overestimated **probability estimates**

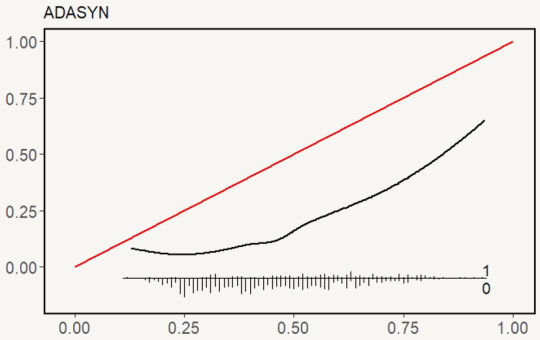
The uncorrected dataset does not lead to overestimation, unlike rebalancing techniques

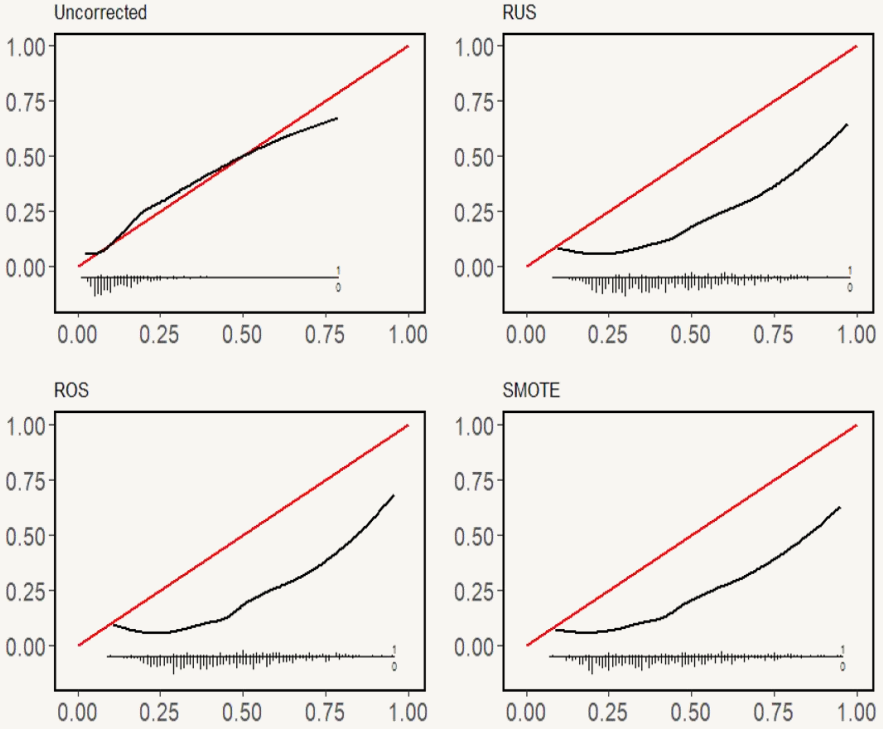


04 – Comparison

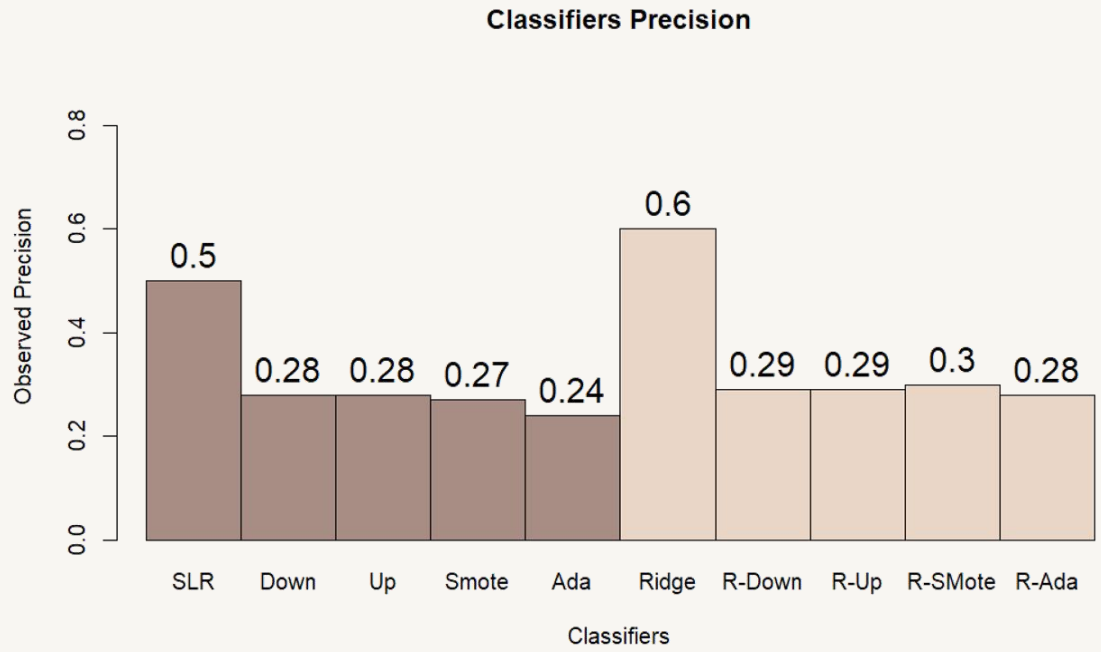
Our experiments, on coronary heart disease dataset, confirm the previous hypothesis.

As for the other rebalancing techniques, **ADASYN** also reports poor results in terms of calibration





05 – Performance analysis



05 – Performance analysis

Probability of being sick
$P(C)$

Sensitivity
$P(+ \mid C) = TP/TP+FN$

Specificity
$P(- \mid C^C) = TN/TN+FP$

Precision
$P(C \mid +) = P(+|C)P(C)/P(+)$

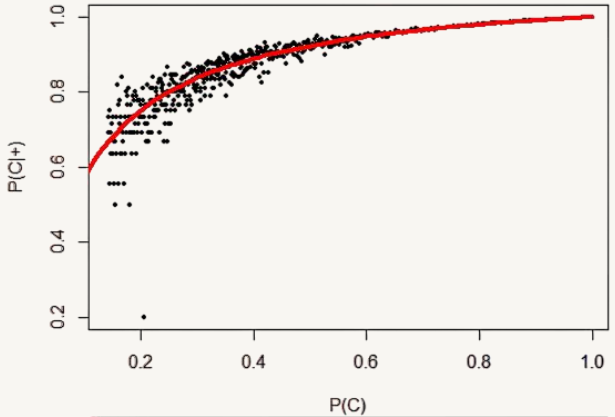
Accuracy
$P(\hat{Y} = Y) = P(+ \mid C)P(C)+P(- \mid C^C)(1-P(C))$

We analyzed the **precision** and **accuracy** as the probability of being sick varies.

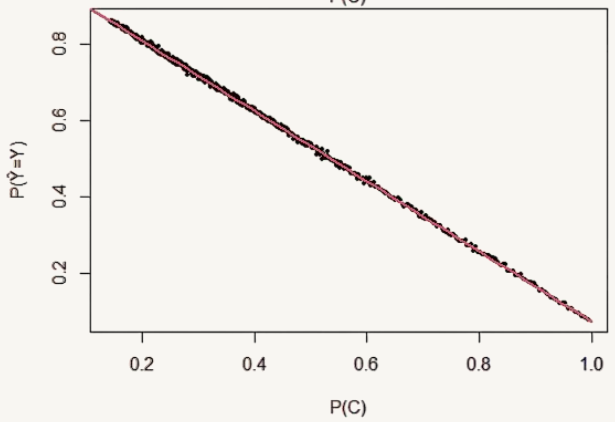
The model used for prediction was a Ridge classifier

05 – Performance analysis

The x axis represent the **percentage of positive** class in the test set.
The y axis represent the **precision** obtained in the different rebalancing percentage
The **red curve** represent the theoretical precision



The x axis represent the **percentage of positive** class in the test set.
The y axis represent the **accuracy** obtained in the different rebalancing percentage
The **red line** represent the theoretical accuracy



06 – Conclusion

Model	Sensitivity (0.5)	Specificity (0.5)	Sensitivity (0.18)	Specificity (0.18)
SLR	0.05	0.99	0.56	0.74
Up	0.62	0.68	0.99	0.10
Down	0.68	0.65	0.98	0.08
Smote	0.65	0.64	0.95	0.22
RIDGE	0.05	0.99	0.55	0.77
Up	0.69	0.65	0.98	0.03
Down	0.71	0.65	0.98	0.02
Smote	0.62	0.71	0.98	0.07

06 – Conclusions

- Our work confirmed the **two hypothesis** advanced in the study:

  (1) Rebalancing techniques distort model calibration

  (2) Using the "imbalance ratio" probability threshold & imbalance correction methods have the same impact on sensitivity and specificity

- The **clinical utility** of the classifier was studied as a function of the risk of overtreatment.

  It emphasized that inaccurate model decisions could lead to unjustified overtreatment.

- It was also shown **how precision and accuracy could vary** according to the probability of being really sick

References

[1] Ruben van den Goorbergh1, Maarten van Smeden 1, Dirk Timmerman2,3, and Ben Van Calster 2,4,5. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. From Journal of the American Medical Informatics Association, 29(9), 2022, 1525–1534 https://doi.org/10.1093/jamia/ocac093 Advance Access Publication Date: 10 June 2022 Research and Applications.

[2] Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi (2015). When is Undersampling Effective in Unbalanced Classification Tasks? ECML/PKDD (1) 200–215. Lecture Notes in Computer Science, volume 9284. https://doi.org/10.1007/978-3-319-23528-8 13

[3] https://en.wikipedia.org/wiki/Framingham_Heart_Study

[4] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset