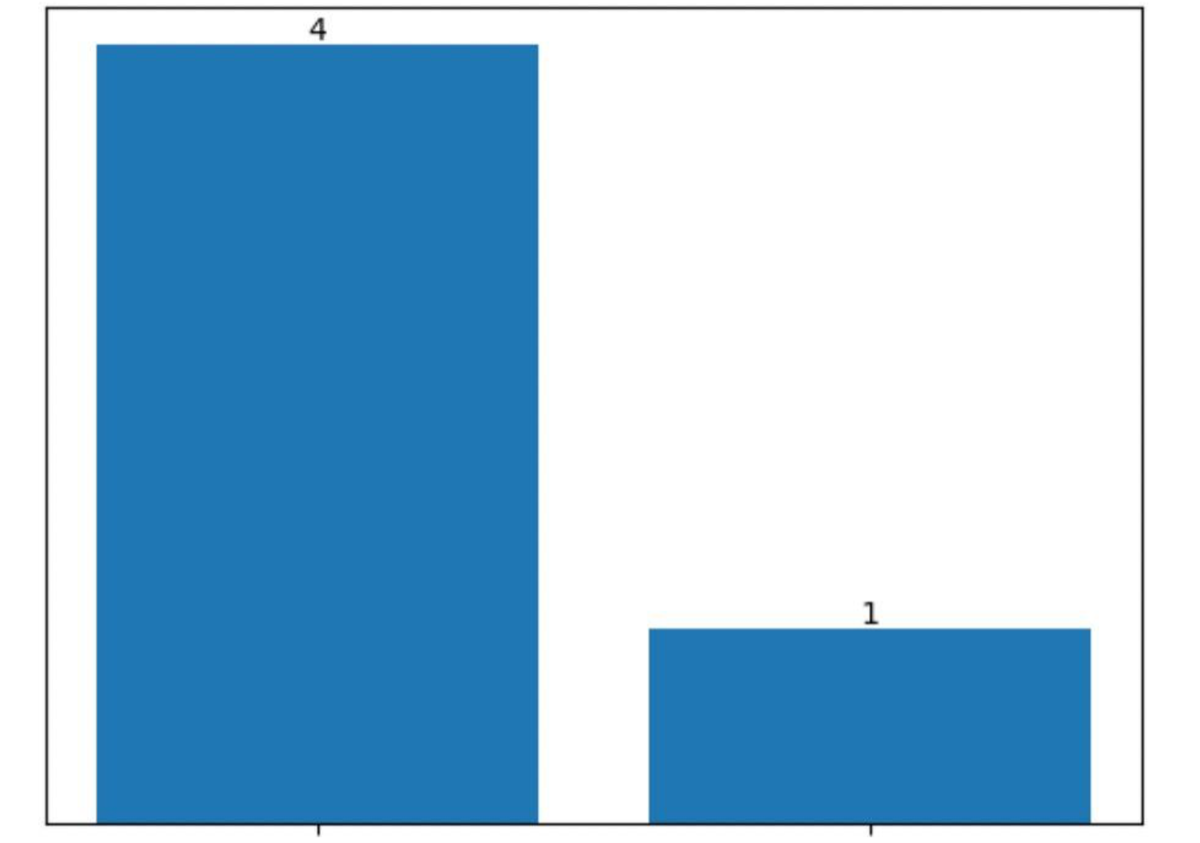
THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Statistics Practitioners' Challenge

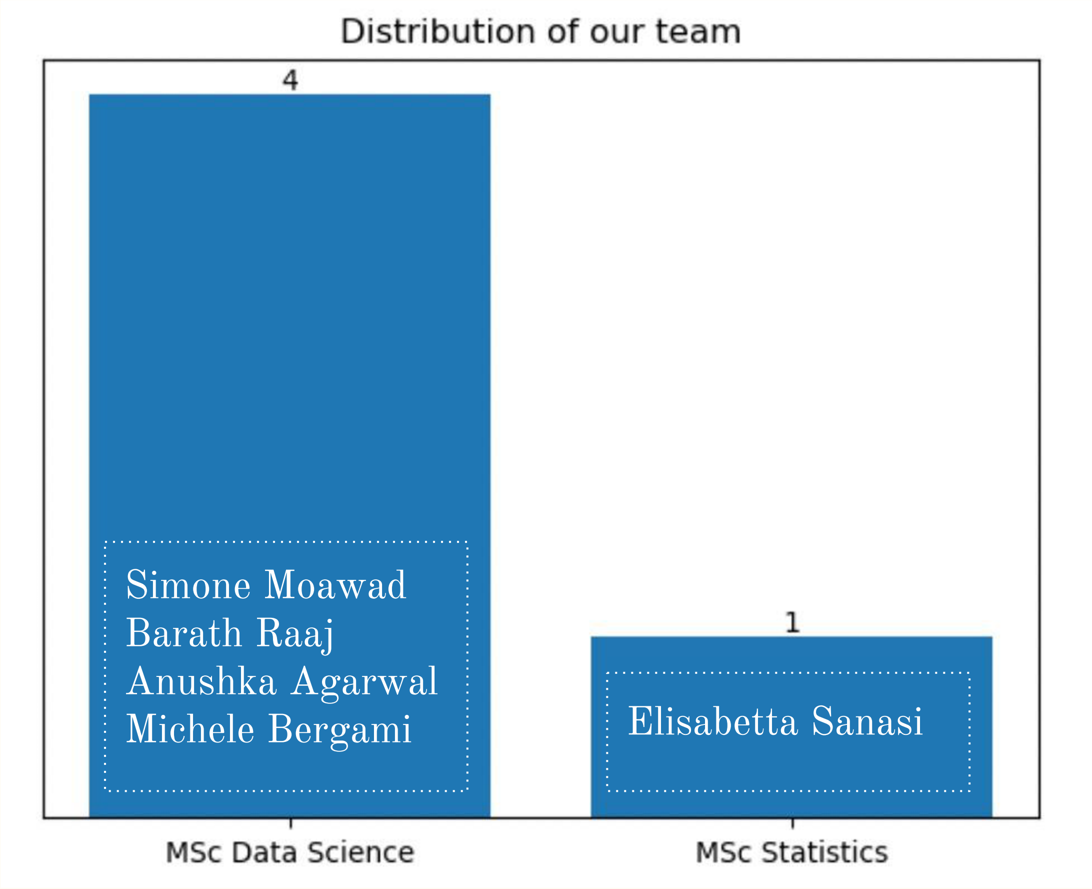
Provided by Allianz

Presentation by "The Rebaggers"

Class imbalances



The team!



Simone Moawad
Barath Raaj
Anushka Agarwal
Michele Bergami

Elisabetta Sanasi

Overview

Introduction

Proposed solution

Resampling techniques

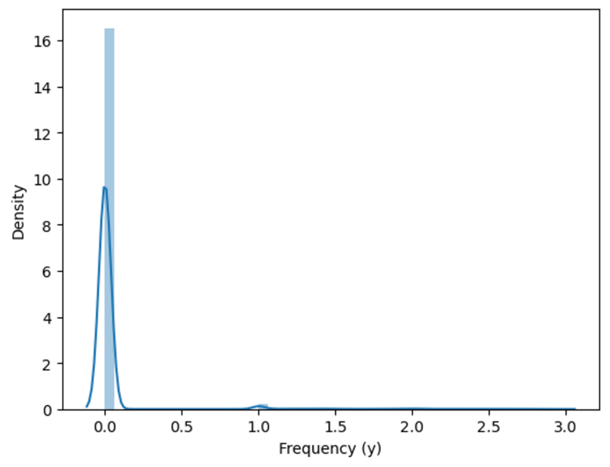
Results

Conclusion

Introduction

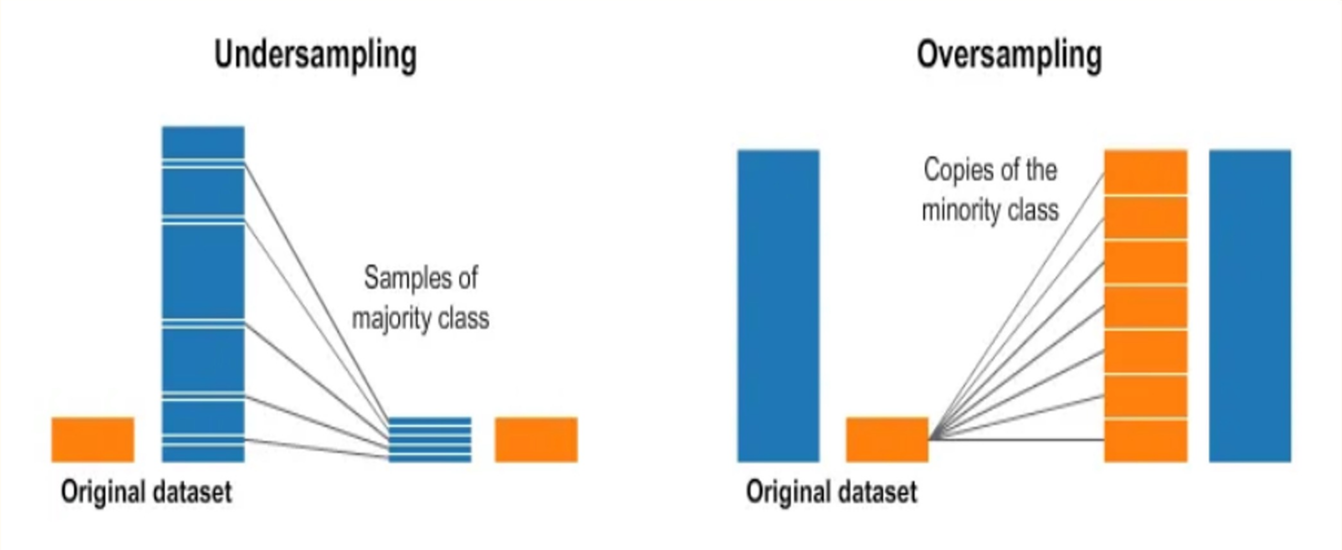
Importance of predicting claim frequency.

- *Fraud detection and prevention*
- *Financial planning and reserving*
- *Pricing of the policies.*



Challenges faced

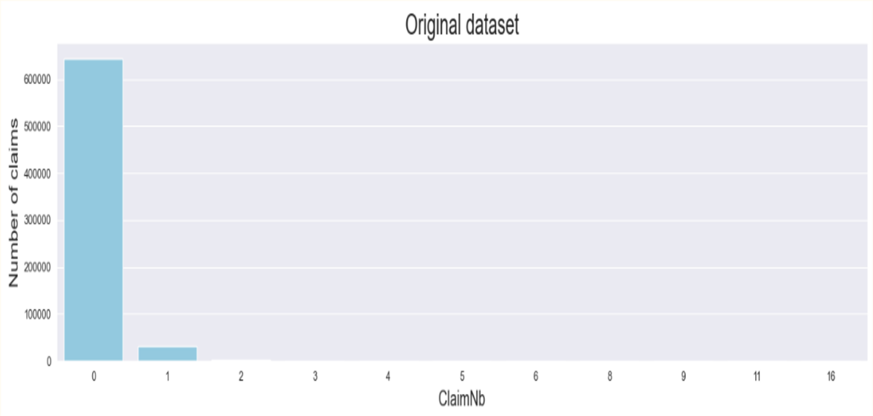
1. Bias and poor performance metrics

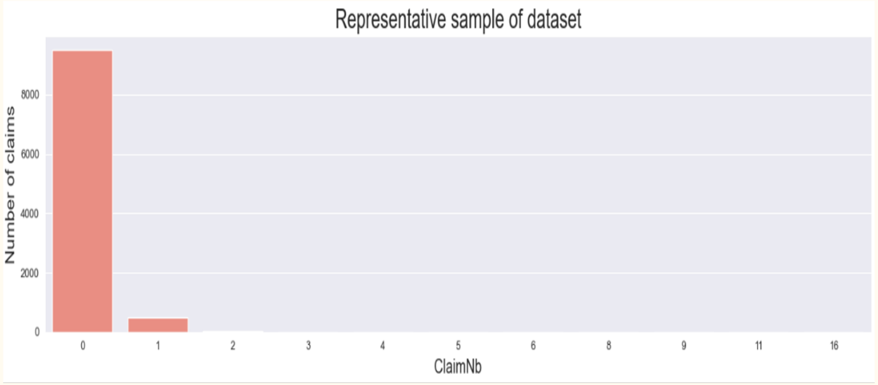


Suggested solution: Application of different resampling techniques to capture the best of both the classes and comparing it within different performance metrics.

Performance metrics - (D2 explained, Mean absolute error, Time taken to train the model, F1 score)

2. Computational complexity and resource constraints





Suggested solution: Representative sampling of the original data to replicate similar results before actually implementing the model on the actual dataset.

Overview

Introduction

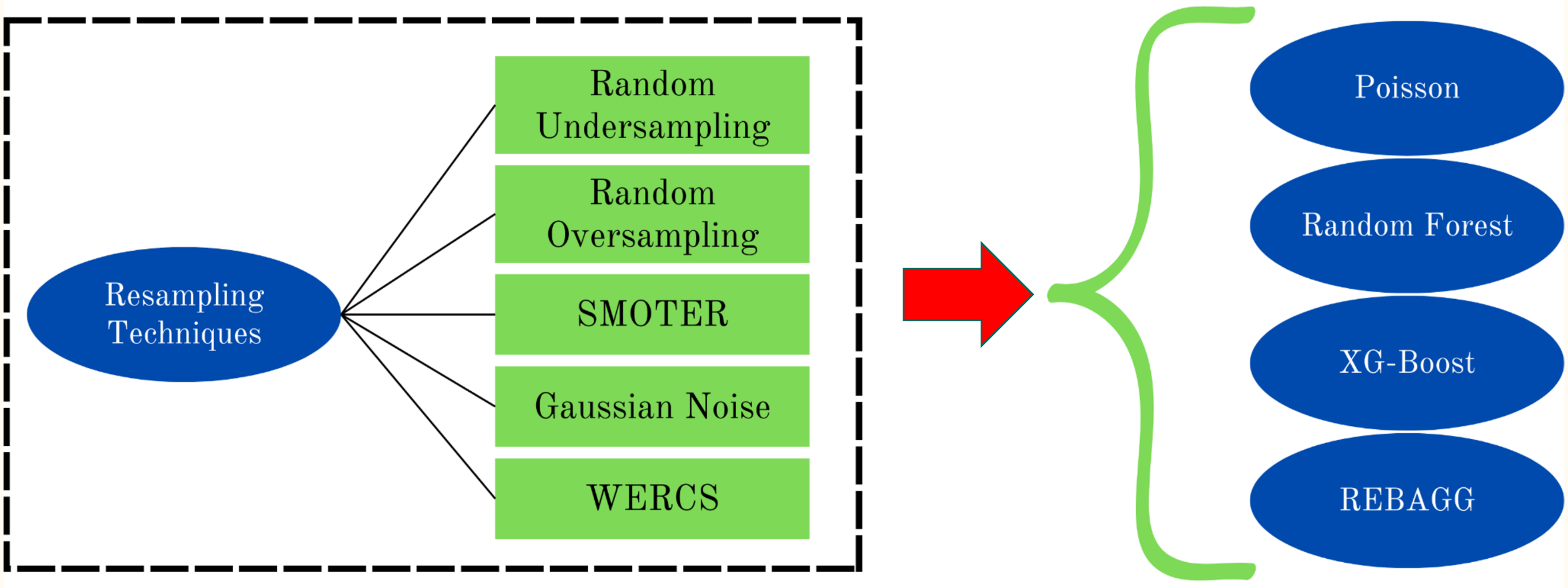
Proposed solution

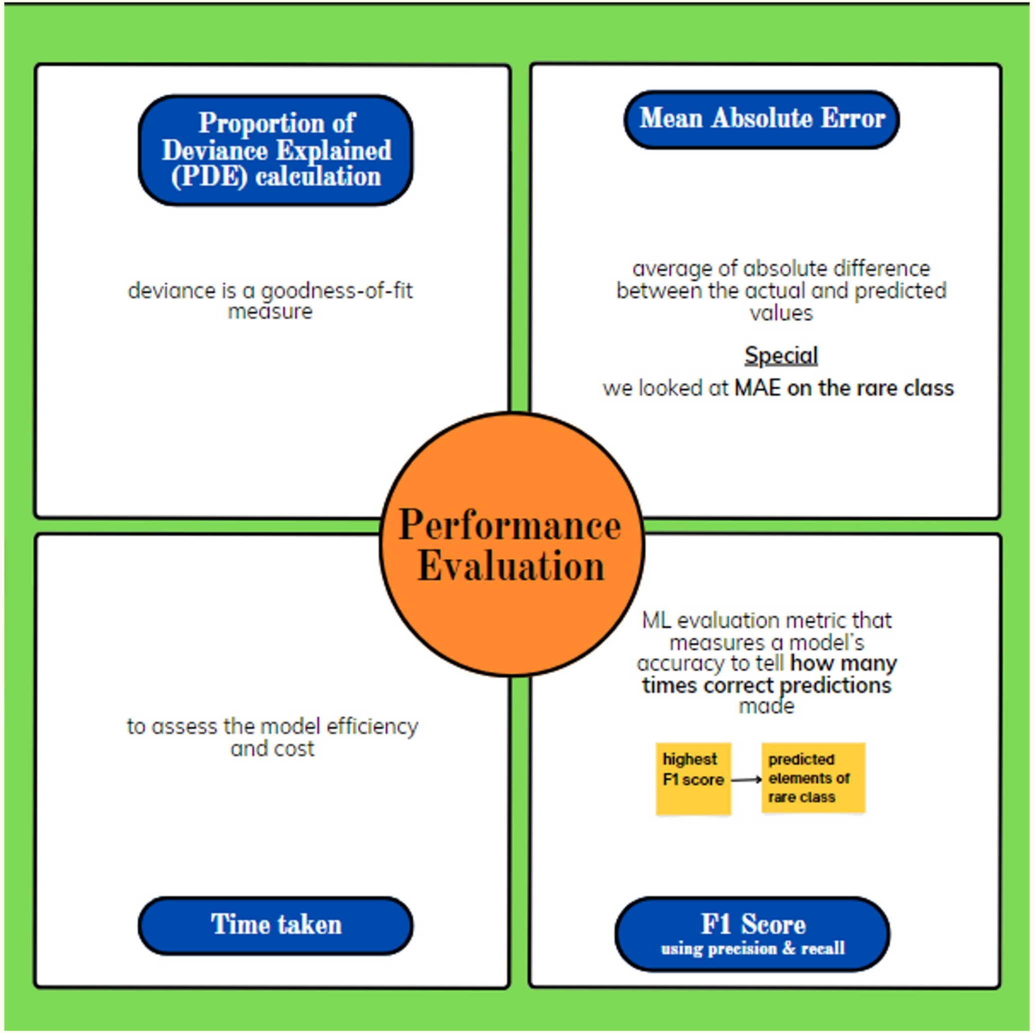
Resampling techniques

Results

Conclusion

Proposed solutions





Performance Evaluation

Proportion of Deviance Explained (PDE) calculation

deviance is a goodness-of-fit measure

Mean Absolute Error

average of absolute difference between the actual and predicted values

<u>Special</u>

we looked at **MAE on the rare class**

Time taken

to assess the model efficiency and cost

F1 Score
using precision & recall

ML evaluation metric that measures a model's accuracy to tell **how many times correct predictions** made

highest F1 score → predicted elements of rare class

Overview

Introduction

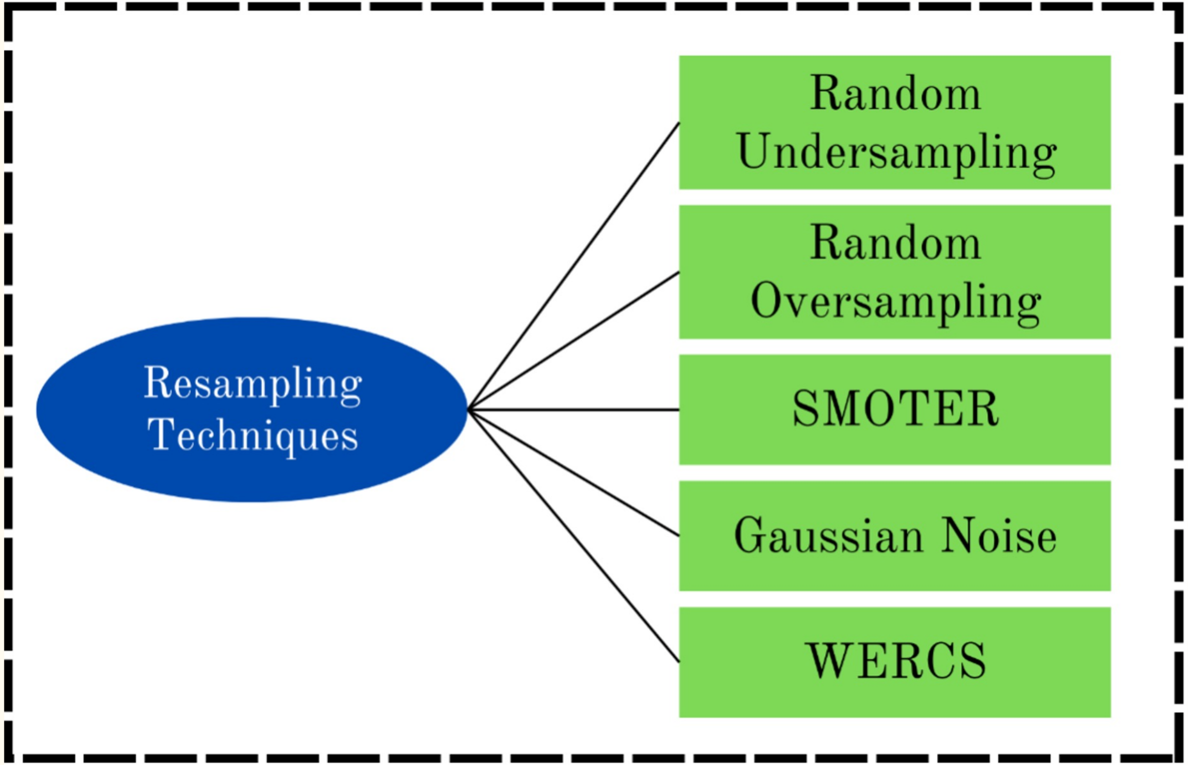
Proposed solution

Resampling techniques

Results

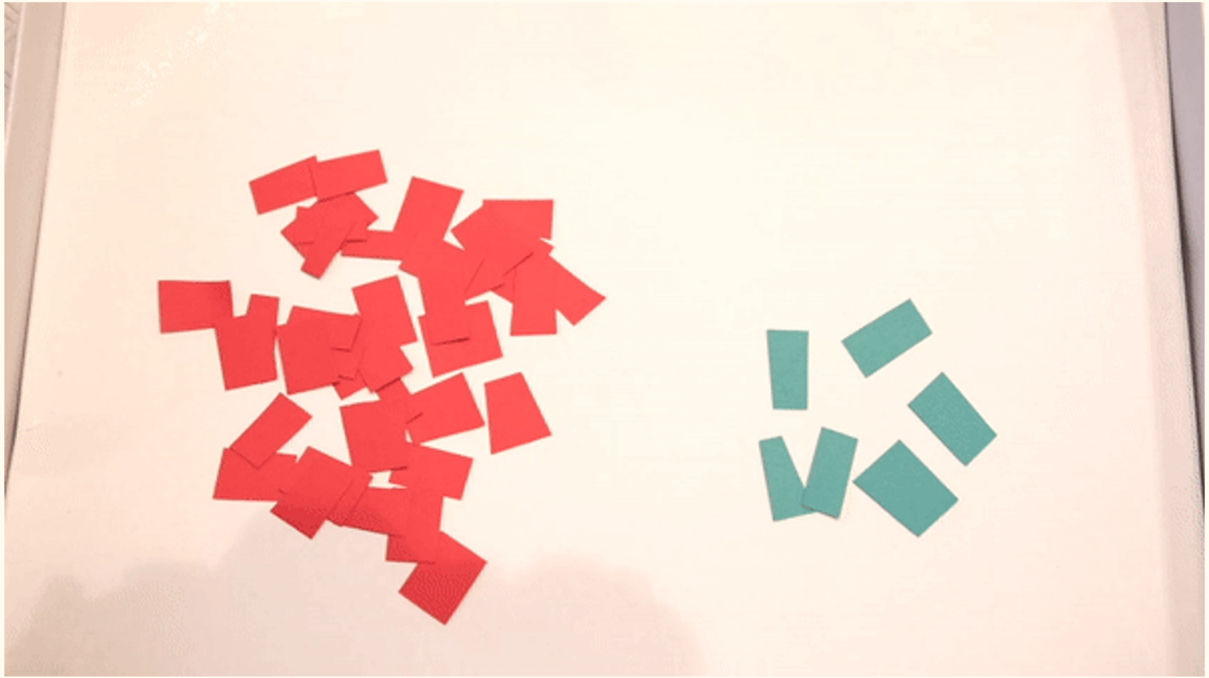
Conclusion

Resampling techniques



What do these techniques mean?

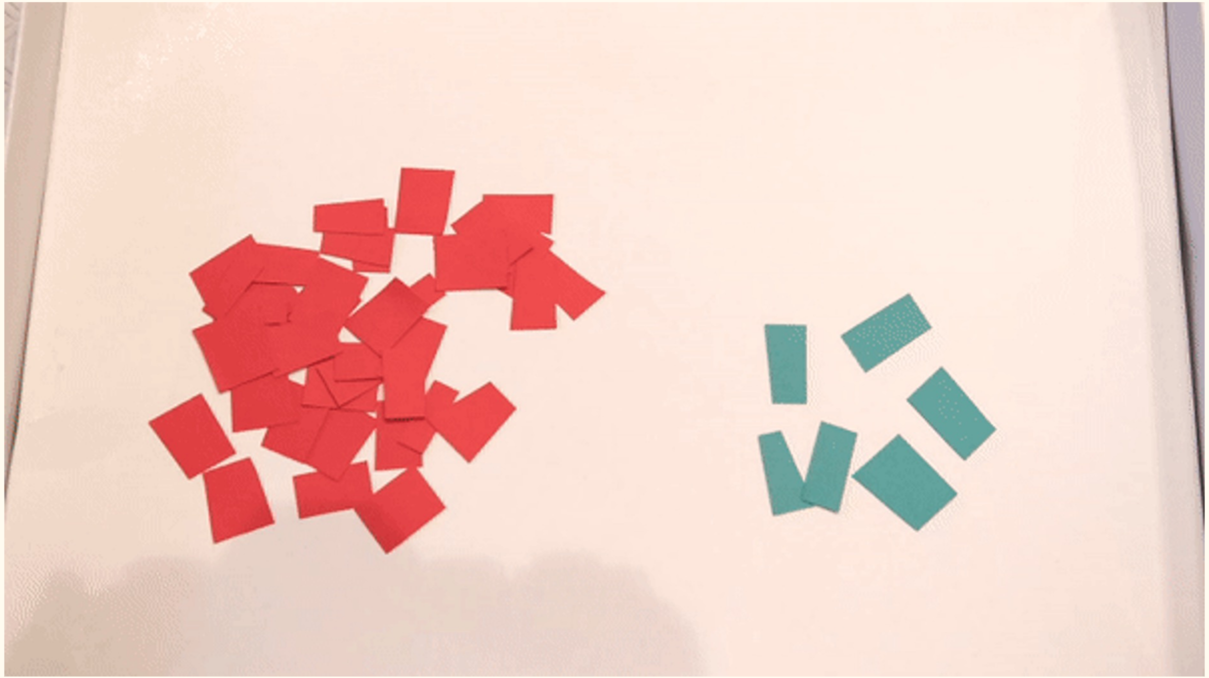
Random undersampling



undersampled
(randomly)

left untouched

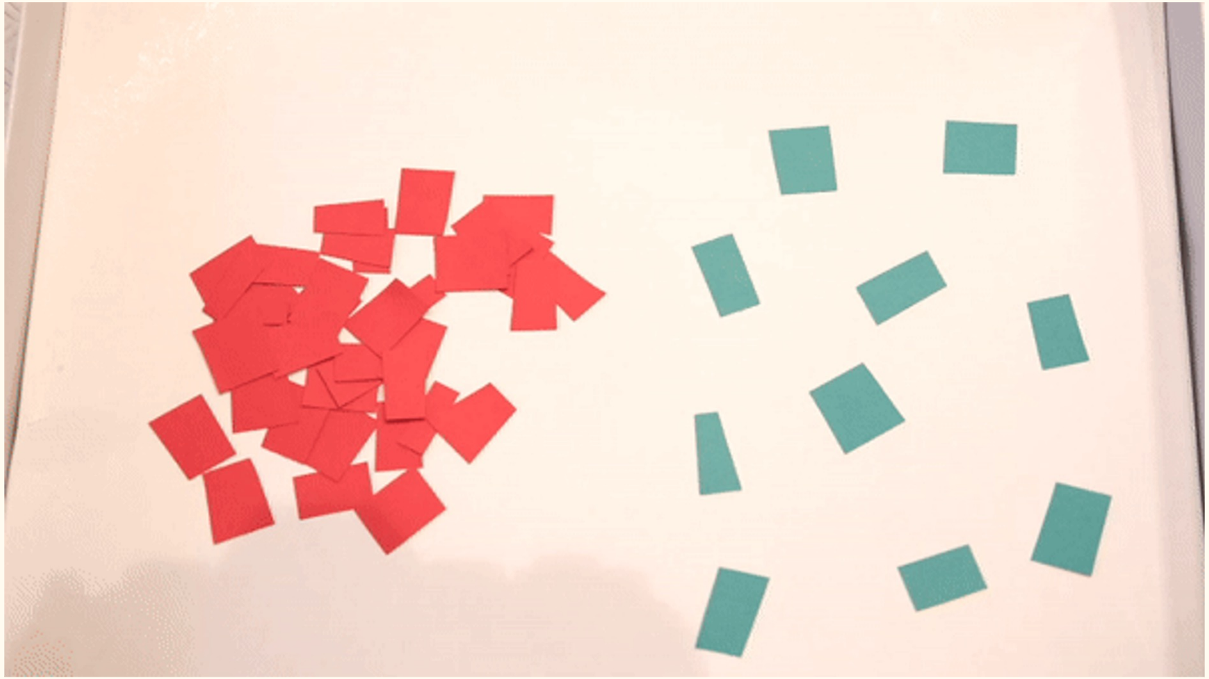
Random oversampling



left untouched

oversampled
(by duplication)

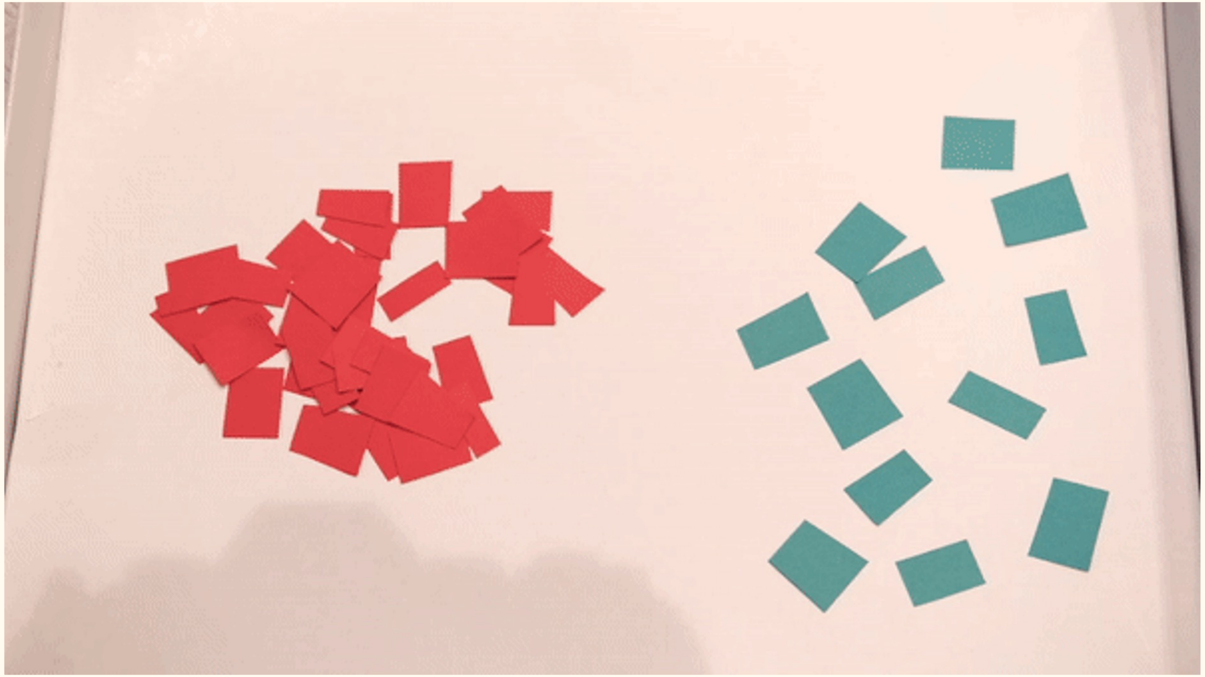
SMOTER



undersampled
(randomly)

oversampled
(using kNN)

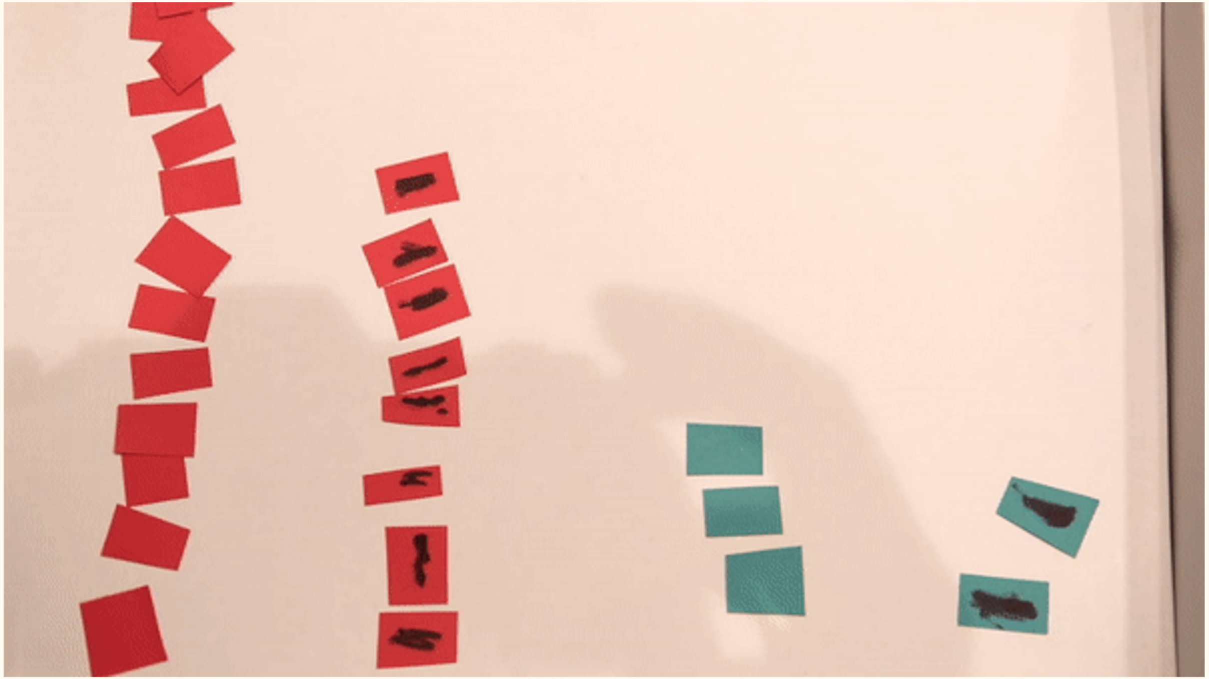
Gaussian noise



undersampled
(randomly)

oversampled
(adding gaussian noise)

WERCS



combination
of the other methods

+

weights
to the data by relevance

Overview

Introduction

Proposed solution

Resampling techniques

Models

Conclusion

The models

Poisson regression

- 💚 Well-suited for count data
- 💔 Strong assumption
- 💔 Linear relationships

Random forest

- 💚 Robust method against overfitting
- 💔 Computationally complex

XGboost

- 💚 Efficient and scalable
- 💔 Tuning of hyperparameters

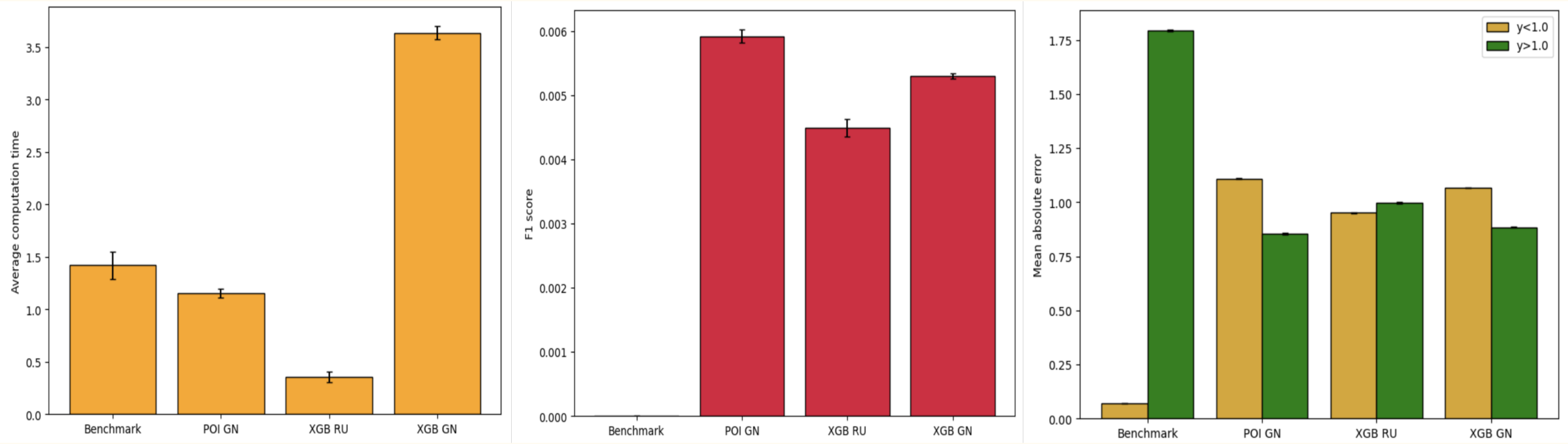
REBAGG

- 💚 Stability and accuracy
- 💔 Computational resources and time for training

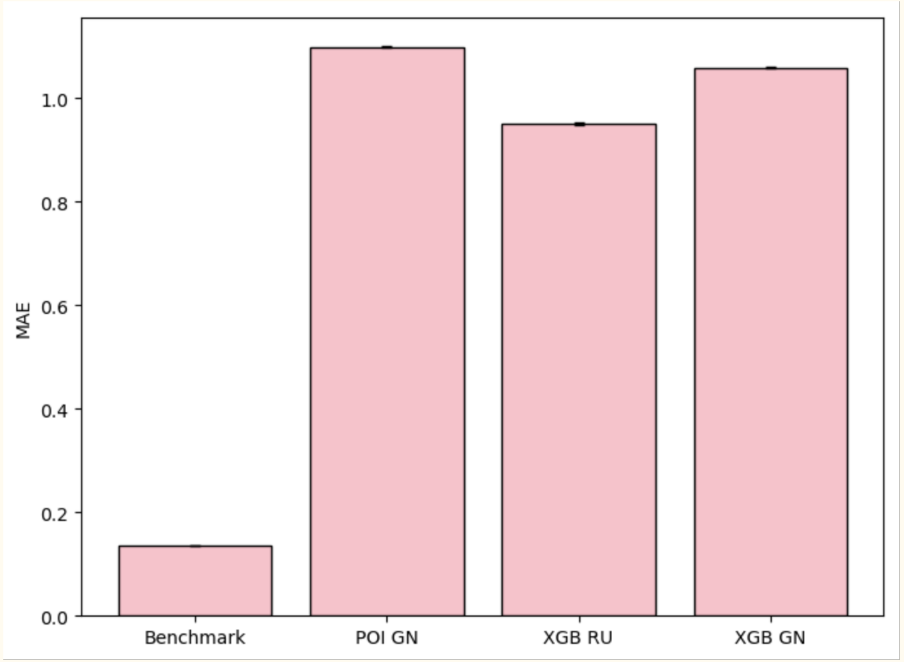
Which is the best model?

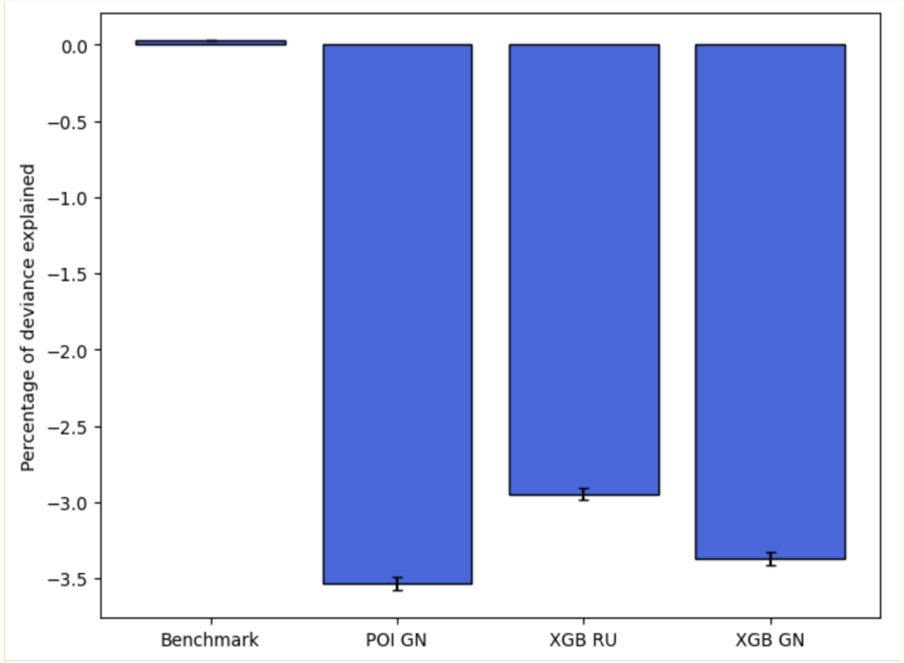
Every model!

Apart from WERCS + Random Forest, that’s terrible :(

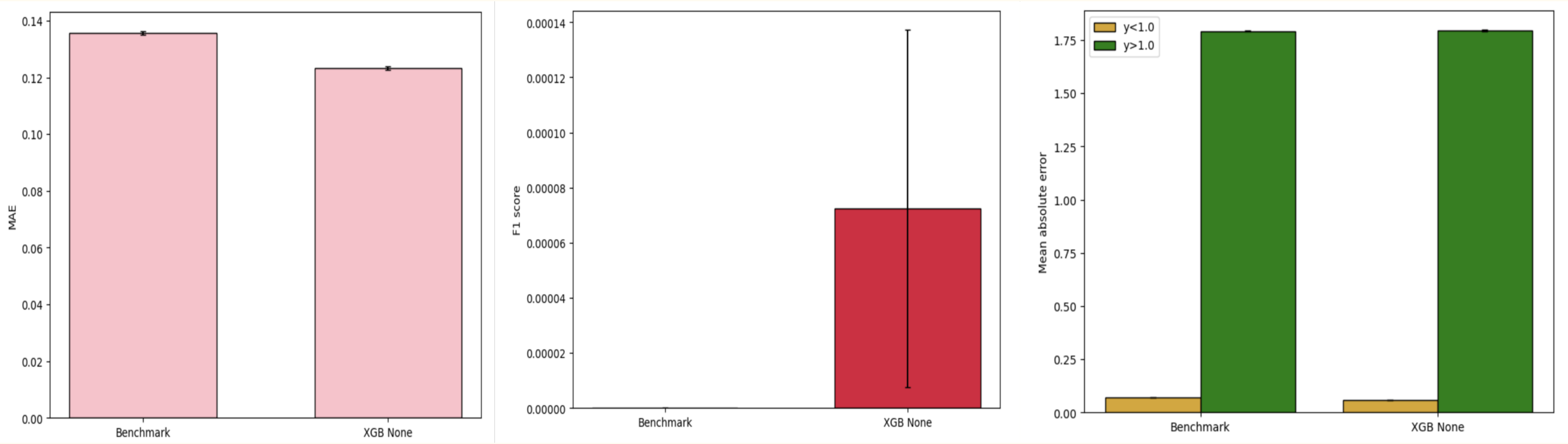


Performance on the rare class

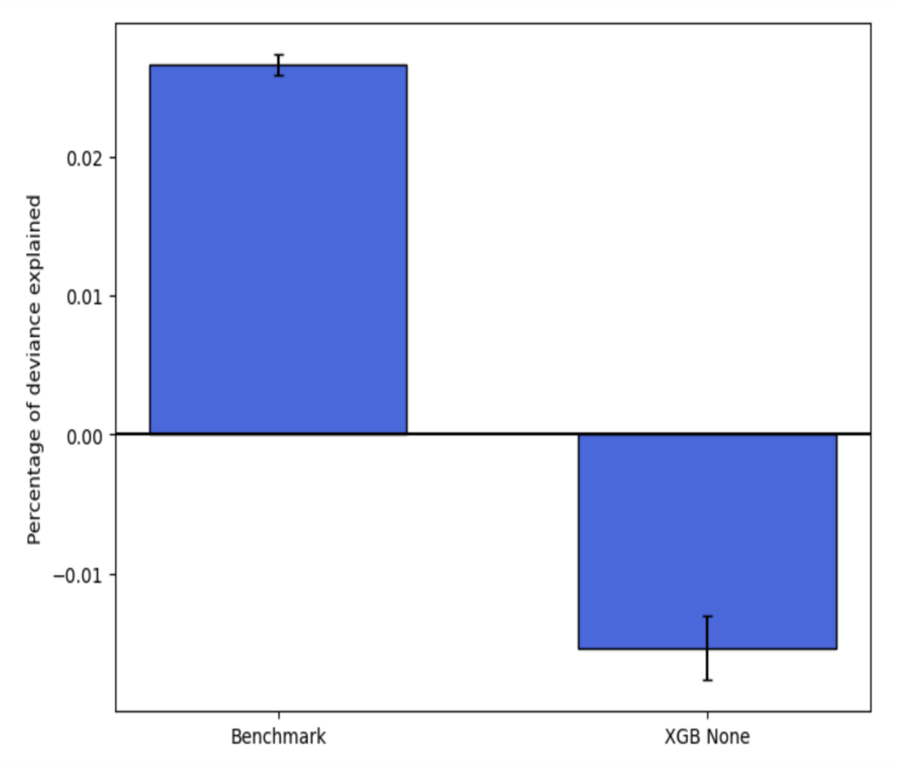


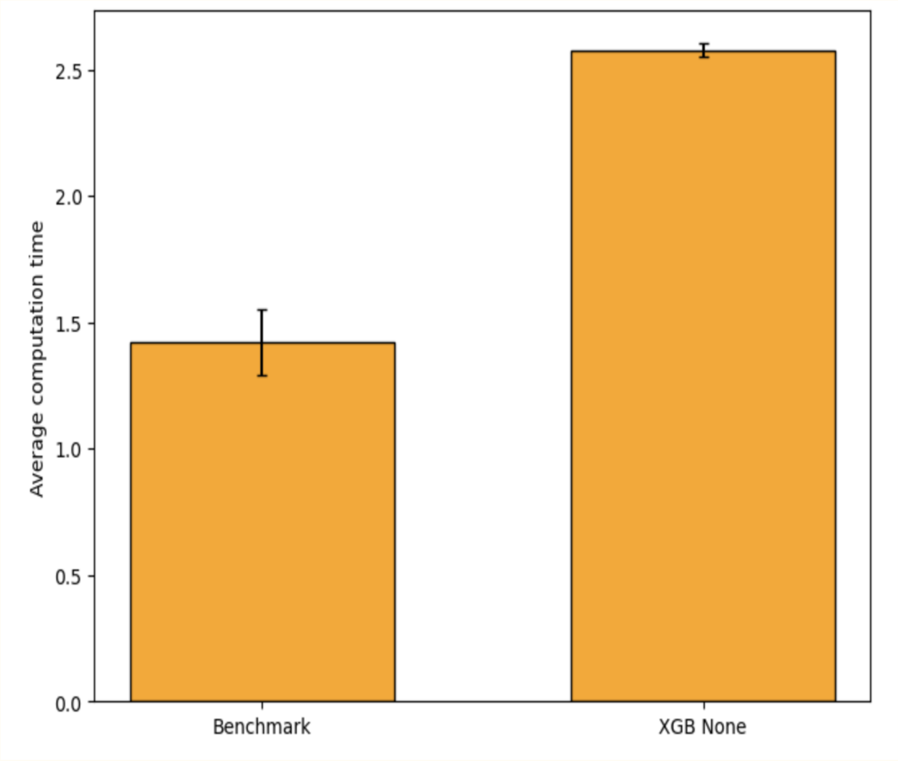


Performance on the rare class

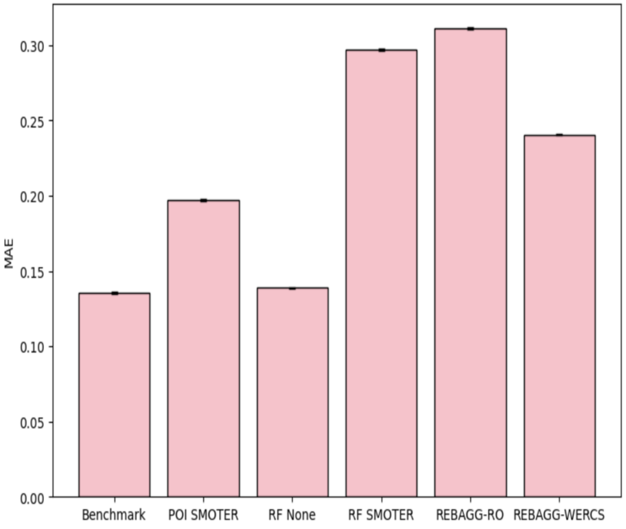


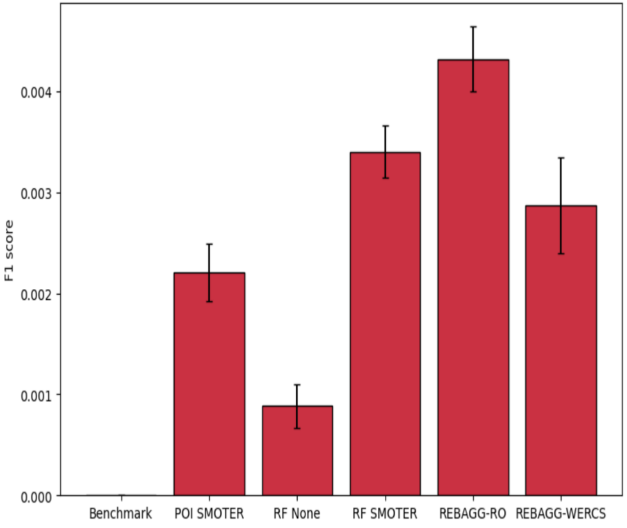
General performance

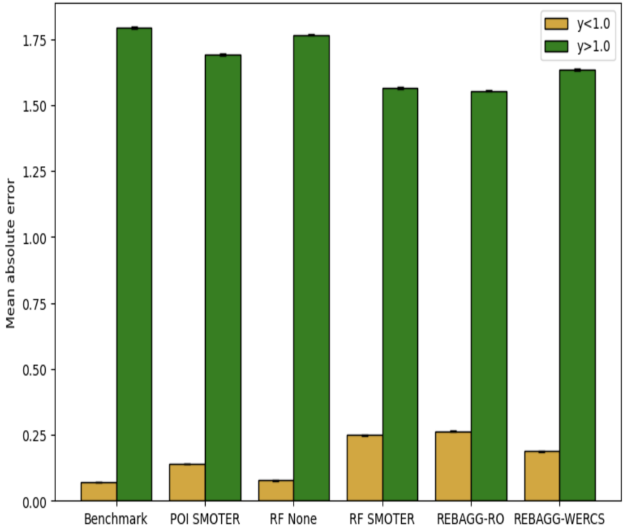




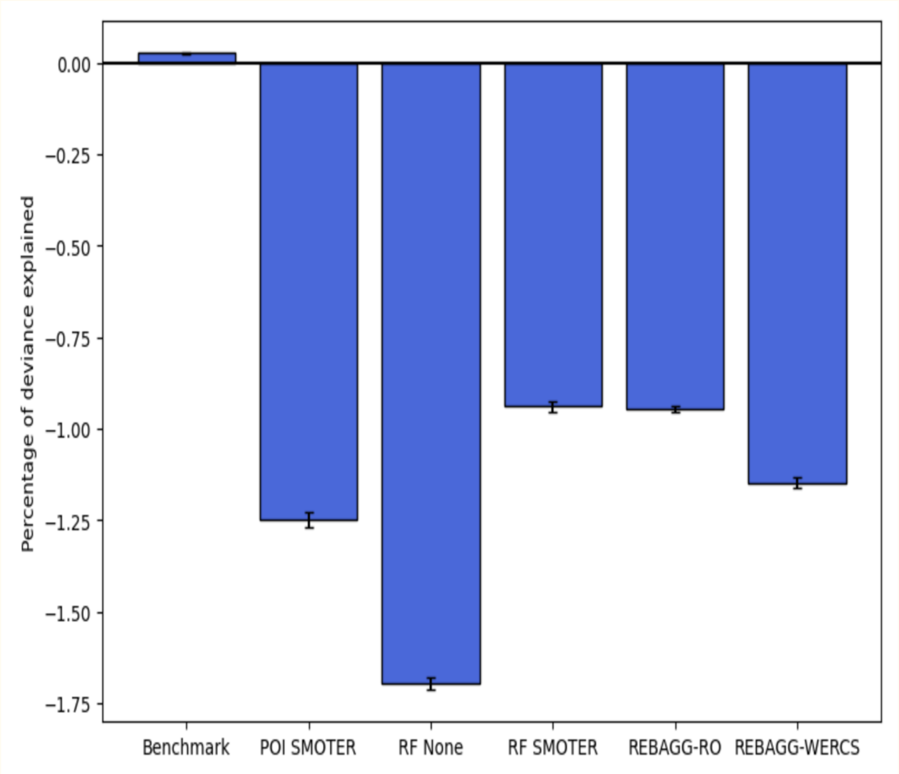
General performance

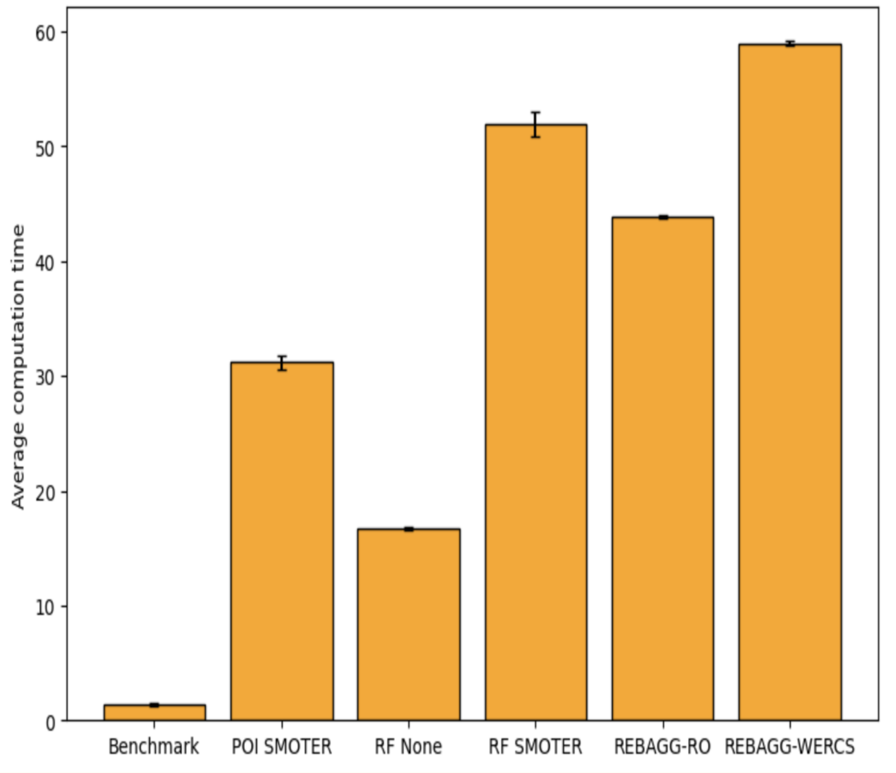






The compromise





The compromise

	PDE	F1 Score	MAE on rare class	MAE	Time
General performance					
Benchmark	0.02660	0.00000	1.79269	0.13565	1.42199
XGB None	-0.01537	0.00007	1.79514	0.12326	2.57665
Performance on the rare class					
Benchmark	0.02660	0.00000	1.79269	0.13565	1.42199
POI GN	-3.53659	0.00592	0.85354	1.09977	1.15724
XGB RU	-2.94768	0.00449	0.99613	0.95204	0.35723
XGB GN	-3.36994	0.00531	0.88476	1.05970	3.63604
Compromise					
Benchmark	0.02660	0.00000	1.79269	0.13565	1.42199
POI SMOTER	-1.24956	0.00220	1.69023	0.19723	31.18506
RF None	-1.69685	0.00088	1.76638	0.13885	16.67995
RF SMOTER	-0.93917	0.00340	1.56522	0.29707	51.94275
REBAGG-RO	-0.94582	0.00432	1.55349	0.31093	43.82252
REBAGG-WERCS	-1.14715	0.00287	1.63469	0.24052	58.96560

The results above are obtained on a device with Apple M1 Chip (8-core CPU, 7-core GPu and 16-core Neural Engine)

Overview

Introduction

Proposed solution

Resampling techniques

Results

Conclusion

Pros

Differentiation of strategies with improvements on MAE in general and on the rare class

Overall computational intensity of selected models that is comparable to the benchmark

Cons

Very low PDE, probably due to the resampling itself

There is not a unique model capable of addressing this model without downsides

Further analysis

Gridsearch on models hyperparameters

Simultaneous gridsearch of resampling techniques and model hyperparameters to possibly achieve an optimal model

Thank you for your attention!