The Effect of Balancing Methods on Model Behavior in Imbalanced Classification Problems

Adrian Stańdo ADRIAN.STANDO.STUD@PW.EDU.PL
Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland

Mustafa Cavus MUSTAFACAVUS@ESKISEHIR.EDU.TR
Eskisehir Technical University, Department of Statistics, Eskisehir, Turkey

Przemysław Biecek PRZEMYSLAW.BIECEK@PW.EDU.PL
Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland

Editors: Nuno Moniz, Paula Branco, Luis Torgo, Nathalie Japkowicz, Michał Wozniak and Shuo Wang.

Abstract

Imbalanced data poses a significant challenge in classification as model performance is affected by insufficient learning from minority classes. Balancing methods are often used to address this problem. However, such techniques can lead to problems such as overfitting or loss of information. This study addresses a more challenging aspect of balancing methods - their impact on model behavior. To capture these changes, Explainable Artificial Intelligence tools are used to compare models trained on datasets before and after balancing. In addition to the Variable Importance method, this study uses Partial Dependence and Accumulated Local Effects profiles. Real and simulated datasets are tested, and an open-source Python package `edgaro` is developed to facilitate this analysis. The results obtained show significant changes in model behavior due to balancing methods, which can lead to biased models toward a balanced distribution. These findings confirm that balancing analysis should go beyond model performance comparisons to achieve higher reliability of machine learning models. Therefore, we propose a new method `performance gain plot` for informed data balancing strategy to make an optimal selection of balancing method by analyzing the measure of change in model behavior versus performance gain.

Keywords: Imbalanced learning, Explainable artificial intelligence, Data balancing, Model behavior change, Performance gain plot

1. Introduction

Classification is one of the most common machine learning (ML) tasks, providing solutions in a wide variety of fields. It frequently involves imbalanced target variables, where there is only one class of particular importance, but there are much fewer data examples available for that class than for the other. This is very common in real-world applications such as credit score prediction, heart attack risk assessment, and fraud detection. However, it can be challenging to train models on such data because many ML models assume a uniform distribution of target variables. If this is not satisfied, the algorithms may lose their ability to learn from the data. One of the approaches to deal with this problem is to apply balancing methods. They are based on undersampling and oversampling, or they combine both approaches.

Although there are numerous balancing methods proposed in the literature, none of them is universally superior. Each has its advantages and disadvantages, and the most appropriate one depends on the specific characteristics of the dataset and the task at hand. For example, oversampling techniques cause excessive learning of the model, which can lead to overfitting, while undersampling methods cause loss of information. Even though such problems, posed by many methods, have been examined from various perspectives, very few studies have examined how they affect how the models behave.

The field that focuses on the study of model behavior is Explainable Artificial Intelligence (XAI). It provides tools that help make the decisions made by models comprehensible and transparent to humans. Consequently, it is possible to understand how balancing methods change the predictions made by the model. This can be done by examining and measuring the extent of changes in the explanations of models trained on the original and balanced datasets. So far, the researchers have only examined these changes by comparing the variable importance (VI) of the models. VI tools provide information about the order of importance of the variables in the model but do not provide any information about the change in the relationships between the variables. In this study, we investigate the effects of the balancing methods on the model behavior using the partial dependence profiles, which determine the relationships between the response variable and the explanatory variables. In addition, we also use accumulated local effect profiles, which are more robust to correlated features and may provide a more accurate representation of model behavior. To measure the extent of change in model explanations, we developed a novel metric called SDD. We used it to compare models trained with logistic regression, random forest, and gradient boosting algorithms on both simulated and real unbalanced datasets. This research uses two different types of data to provide a more controlled assessment of changes in model behavior. This is supported by the fact that after applying balancing methods to real-world datasets, it is difficult to determine which one represents the true ground truth because the model structure has changed. To address this issue, we also perform simulations using synthetic datasets where the ground truth is known. Furthermore, we propose the performance gain plot, which can be used to select the optimal balancing method to solve the dilemma arising from the negative effects of balancing methods on model behavior and improve model prediction performance. In addition, to facilitate the evaluation of different balancing methods and XAI tools, we have developed a Python package that provides a unified interface for data preprocessing, model training, and XAI analysis. The package includes implementations of several popular balancing methods and XAI tools, as well as customized evaluation metrics to measure the impact of balancing methods on model behavior.

The main contributions of this paper are as follows: (1) to investigate the effects of balancing methods on model behavior, (2) to propose a measure to quantify changes in model behavior, (3) to create benchmark datasets with imbalanced class distributions, (4) to propose the performance gain plot for optimal balancing method selection regarding the performance gain versus the model behavior change, and (5) to develop a Python package that simplifies the workflow of using balancing methods, training models, and applying XAI tools. In the remainder of this paper, we first discuss the related works in Sec. 2, then we present the XAI tools used in the experiments and the proposed comparison metric in

Sec. 3, describe our experiments conducted on simulated and real datasets in Sec. 4, and discuss the results and conclusions in Sec. 5.

2. Related Works

The focus of this paper is to explore how XAI tools offer a new perspective on the problem of imbalanced learning. In this section, we provide an overview of existing methods for addressing the issue of imbalanced learning and a summary of research on their impact on model behavior. In addition, we emphasize the distinctive contributions of our study compared to similar work in the field.

In many domains where binary classification is applied, the class of interest is extremely rare. Classification models tend to favor the majority class in such cases, leading to bias. This bias results in a higher frequency of misclassification of minority class examples. The problem of bias towards the majority class has been addressed through several proposed methods, which can be divided into two groups: algorithmic-level methods (Gu et al., 2022; Li et al., 2022), which aim to develop better algorithms, and data-level methods (Chawla et al., 2002; He et al., 2008), which involve transforming the original dataset to balance it.

There are many studies focused on how data balancing influences and changes the performance of trained ML models (Ortega Vázquez et al., 2023; Gu et al., 2022). However, there have not been many attempts to investigate how they affect the model's behavior. Patil et al. (2020) investigated the changes in the order of importance of the variables in the model after applying the balancing methods. They studied whether the balancing technique SMOTE changes the correlations between features. The experiment was conducted only on one highly imbalanced dataset. The results show that this algorithm was successful not only in eliminating imbalance but also in preserving the original correlations. The authors then applied a few XAI methods to extract the explanations of the model trained on oversampled data. However, they emphasized that it could be done only because the feature correlations remained unchanged. Alarab and Prakoonwit (2022) sought an answer to a similar question in an application on blockchain data. They compared the explanations of models trained on the dataset after applying different balancing methods. The feature importance is used and compared with a statistical test. The experiments were conducted on two datasets using different variants of the SMOTE algorithm. The results show that one of the methods changed the feature importance in both cases. However, these studies have two main limitations: (1) they do not provide enough comprehensive information about the change in model behavior as they only use feature importance, and (2) their results cannot be generalized because they are based on only one or two datasets. Moreover, Saarela and Jauhiainen (2021) showed that the most important features differ depending on the variable importance technique used. They suggested using a combination of the explanation techniques could provide more consistent results.

In this paper, we aim to investigate the effects of balancing methods on model behavior, which is firstly mentioned in Cavus and Biecek (2022) and addressed in more detail in Stańdo (2023). To do so, we propose a new metric based on the differences between the partial dependence profiles or the accumulated local effect profiles, since it is not possible to directly measure the model behavior over the PDP and ALE profiles by using the existing

metrics proposed in Schwalbe and Finzel (2023); Visani et al. (2022); Roy et al. (2022); Zhang et al. (2022); Agarwal et al. (2022).

3. Methodology

This section presents the XAI tools used and the proposed metric for measuring the change in model behavior.

3.1. Partial dependence profile

The partial dependence profile (PDP) is introduced in Friedman (2001). Let X be the data set and X^j be any variable in the data set. The PDP is a function of the observation z for a model f and a variable j defined as $PDP(f, j, z) = E_{X^{-j}}[f(X^{j|=z})]$. In other words, the PDP value for the j-th column in the observation z is an average prediction of model f when values in the j-th column are set to z. In practice, however, we do not usually know the distribution of X^{-j} (Biecek and Burzykowski, 2021). Therefore, it is estimated using $\widehat{PDP}(f, j, z) = \frac{1}{n}\sum_{i=1}^{n} f(X_i^{j|=z})$.

3.2. Accumulated local effects

PDP may provide explanations that can be misleading if explanatory features are correlated. Therefore, the Accumulated Local Effects (ALE) profiles are proposed (Apley and Zhu, 2020). It is a noteworthy alternative to PDP because both produce the functions as outputs, but ALE is unbiased. ALE for a model f and a variable j is a function of observation z defined as follows:

$$ALE(f, j, z) = \int_{z_0}^{z} \left(E_{X^{-j}} \left[\frac{\partial f(u)}{\partial u^j}_{u=X^{j|=v}} \right] \right) dv + c. \tag{1}$$

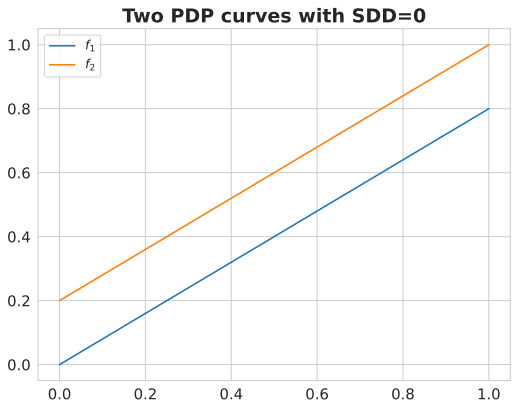
The constant c is selected in such a way that $E_{X^j}[ALE(f, j, X^j)] = 0$ and z_0 is a value close to the lower bound of the support of X^j. In other words, $\frac{\partial f(u)}{\partial u^j}$ describes the local change of the model which is then averaged over the distribution of X^{-j} and integrated over values from z_0 to z. The estimator of ALE is the following:

$$\widehat{ALE}(f, j, z) = \sum_{k=1}^{K} \left(\frac{1}{\sum_l w_l^j(z_k)} \sum_{i=1}^{N} w_i^j(z_k) \left[f\left(x^{j|=z_k}\right) - f\left(X^{j|=z_k-\Delta}\right) \right] \right) + \hat{c}. \tag{2}$$

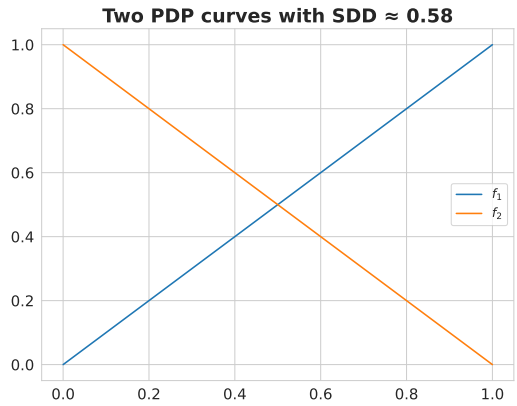
The constant K is the number of intermediate points, $(z_0, z_1, ..., z_K)$ are evenly distributed points in (z_0, z) interval with step $\Delta = (z - z_0)/K$. The weights $w_i^j(z_k)$ represent the distance between z_k and x_i^j. The constant $\hat{c}$ is selected in such a way that $\sum_{i=1}^{N} ALE(f, j, X_i^j) = 0$ (Biecek and Burzykowski, 2021).

3.3. Variable importance

The previous two methods show how the model predictions change when the value of the selected variable changes. The variable importance (VI) tool, however, focuses on creating one explanation for all variables in the model. We use the VI which permutes each column in X multiple times and see how it affects the model performance - the method proposed by Fisher et al. (2019).



(a) Two example PDP curves with $SDD = 0$.



(b) Two example PDP curves with $SDD \approx 0.58$.

Figure 1: Exemplary PDP plots and the SDD values.

3.4. Standard Deviation of the Differences (SDD)

Assume that f_1 and f_2 be two models trained on the same data set, the SDD metric:

$$SDD(h_1, h_2, j, k) = sd[h_1(x_i) - h_2(x_i)_{i=1,2,...,k}] \tag{3}$$

where $h_1(z) = PDP(f_1, j, z)$ and $h_2(z) = PDP(f_2, j, z)$ are the values of the profiles for the j-th variable at point z. In the above definition, if $h_1(z) = h_2(z), \forall z$, which means that the profiles are equal, the metric value $SDD(h_1, h_2, j, k) = 0$. Similarly, if $h_1(z) = h_2(z) + c, \exists c \forall z$, which means that the profiles are parallel and $SDD(h_1, h_2, j, k) = 0$. This behavior is expected as the metric is intended to measure the changes in the shape of the curves, not the vertical offset. This is because the position of the PDP curve depends on the accuracy of the model, but only the changes in shape indicate changes in behavior.

On the other hand, consider $(x_1, x_2, ..., x_{101}) = (\frac{0}{100}, \frac{1}{100}, ..., \frac{100}{100})$, $h_1(z) = z$, and $h_2(z) = 1 - z$. In such a situation, $SDD(h_1, h_2, j, 101) \approx 0.58$. As can be seen in Figure 1, the curves that behave differently have a high value of SDD. It compares two models for one variable. Additionally, the SDD values can be aggregated to the averaged SDD ($ASDD$) values to compare the behavior of models with respect to all variables. This can be formalized as $ASDD(f_1, f_2, k) = \frac{1}{m} \sum_{j=1}^{m} SDD[PDP(f_1, j, *), PDP(f_2, j, *), j, k]$.

4. Experiments

In this section, we conduct experiments to measure the impact of six balancing methods: Undersampling (Random, Near Miss), Oversampling (Random, SMOTE, Borderline SMOTE), and Hybrid (SMOTETomek) on the behavior of the three models (Logistic Regression, Random Forest, and XGBoost) behaviors on simulated and real imbalanced datasets in terms of SDD of PDP and ALE. The `edgaro` Python package (Explainable imbalanceD learninG compARatOr) is implemented to run the experiments. It is the first to provide a user-friendly interface to balance and train ML models for several datasets arranged in arrays or

nested arrays. It allows using implementations from two major libraries, *scikit-learn* and *imbalanced-learn*. The package also calculates, in the same form of arrays and nested arrays, explanations using the PDP, ALE, or VI method and provides functions to compare them. To ensure that the explanations for each experiment are calculated over the same data, the test dataset extracted before any balancing is used as the background data. By default, the test size is equal to 20%, and the data is split in a stratified fashion (preserving the class distribution in both subsets).

4.1. Simulated Dataset Experiments

We conducted experiments on simulated data, which were generated through simulations that allowed us to control the ground truth (Amiri et al., 2020). We were unsure which model represented the ground truth, as there are changes in the model behavior due to the use of balancing methods. For this purpose, we constructed a comprehensive model framework and followed a simulation design similar to Casalicchio et al. (2019), and we used the following model to simulate the imbalanced datasets:

$$z = \beta_{0i} + 2.9X_1 - 3.7X_2 + 1.2X_3 + \epsilon_j \tag{4}$$

where the binary response variable $Y \sim B(1, p = 1/(1+exp(-z)))$, the explanatory variables $X_1, X_2, X_3 \sim N(0, 1)$ and the error term $\epsilon_j \sim N(0, v_j)$. The simulations are set up as 12 scenarios: β_{0i} takes the values $\{1.5, 2.5, 3.5, 4.5\}$ and the variance of the error term v_j takes the values in $\{1, 2, 3\}$ respectively, in the scenario ij to generate the dataset. The error term ϵ controls the variance of the model prediction, and the parameter β_{0i} controls the imbalance ratio of the target variable in the dataset.

The figure in https://tinyurl.com/baccuracyplot, shows how the balancing methods improve the model performance in terms of balanced accuracy, at the bottom of each panel shows the distribution of the model performance, which is trained on imbalanced datasets as the reference level. As the value of the coefficient β_0 increases, indicating an increase in the imbalance ratio, the performance of the model decreases. Similarly, an increase in the variance of the error term leads to a decrease in the model's performance. When evaluated separately, logistic regression, random forests, and XGBoost exhibit the highest performance, respectively. Among them, increasing the variance of the error term has the largest impact on reducing the performance of the random forest model. When examining the positive effects of balancing methods on model performance, Random Undersampling is the most effective method, followed by all Oversampling methods. The Near Miss method reduces the performance of the random forest and XGBoost models as the imbalance ratio increases.

The SDD values for PDP profiles between models trained on the simulated datasets are presented in Figure 2 (the figure in https://tinyurl.com/alesdd for the ALE profiles). The similarity of SDD values for PDP and ALE indicates that either approach can be used to compare models. The impact of undersampling and other methods on model behavior varies. Undersampling has the smallest effect on the Logistic Regression model but the largest effect on the XGBoost model. Conversely, other methods have a greater impact on the behavior of the logistic regression model and a lesser impact on the random forest model. The effects of these methods become more pronounced as the variance of the model error

Table 1: Proposed benchmarking set

Dataset name	IR	Rows	Columns	Source
spambase	1.54	4601	55	OpenML-100, OpenML-CC18
MagicTelescope	1.84	19020	10	OpenML-100
steel-plates-fault	1.88	1941	13	OpenML-100, OpenML-CC18
qsar-biodeg	1.96	1055	17	OpenML-100, OpenML-CC18
phoneme	2.41	5404	5	OpenML-100
jm1	4.17	10880	17	OpenML-100, OpenML-CC18
SpeedDating	4.63	1048	18	OpenML-100
kc1	5.47	2109	17	OpenML-100, OpenML-CC18
churn	6.07	5000	8	OpenML-CC18
pc4	7.19	1458	12	OpenML-100, OpenML-CC18
pc3	8.77	1563	14	OpenML-100, OpenML-CC18
abalone	9.68	4177	7	imblearn
us_crime	12.29	1994	100	imblearn
yeast_ml8	12.58	2417	103	imblearn
pc1	13.40	1109	17	OpenML-100, OpenML-CC18
ozone-level-8hr	14.84	2534	72	imblearn, OpenML-100, OpenML-CC18
wilt	17.54	4839	5	OpenML-100, OpenML-CC18
wine_quality	25.77	4898	11	imblearn
yeast_me2	28.10	1484	8	imblearn
mammography	42.01	11183	6	imblearn
abalone_19	129.53	4177	7	imblearn

term and the imbalance ratio of predicted values increase. Thus, the negative impact of balancing methods on model behavior increases with higher variance and imbalance ratios in the model predictions.

4.2. Benchmark Datasets

Benchmark datasets are the backbone of large-scale experiments. Their quality is of great importance for generalizing the results obtained in the experiments. Moniz and Cerqueira (2021) and Singh and Vanschoren (2022) proposed benchmark datasets for imbalanced learning. To measure the effect of balancing methods on model behavior in terms of SDD, we propose a new benchmarking set of datasets. For now, SDD works only on continuous variables, therefore, we need to create a new imbalanced benchmark dataset.

The proposed benchmarking set of datasets is made up of three main sources: OpenML-100, OpenML-CC18 (Bischl et al., 2017), and the collection of datasets available in `imblearn` library which was proposed by Ding (2011). The benchmarking set contains only datasets for binary classification tasks that have only continuous columns (categorical and nominal were removed), at least 1000 rows, and an imbalance ratio of at least 1.5. The set is also available via a dedicated class in `edgaro` package. The list of selected datasets and their details are presented in Table 1.

4.3. Real Dataset Experiments

We ran the experiments on the proposed benchmark dataset. Figure 3(a) presents the balanced accuracy values for models on both original and balanced datasets. It is clear

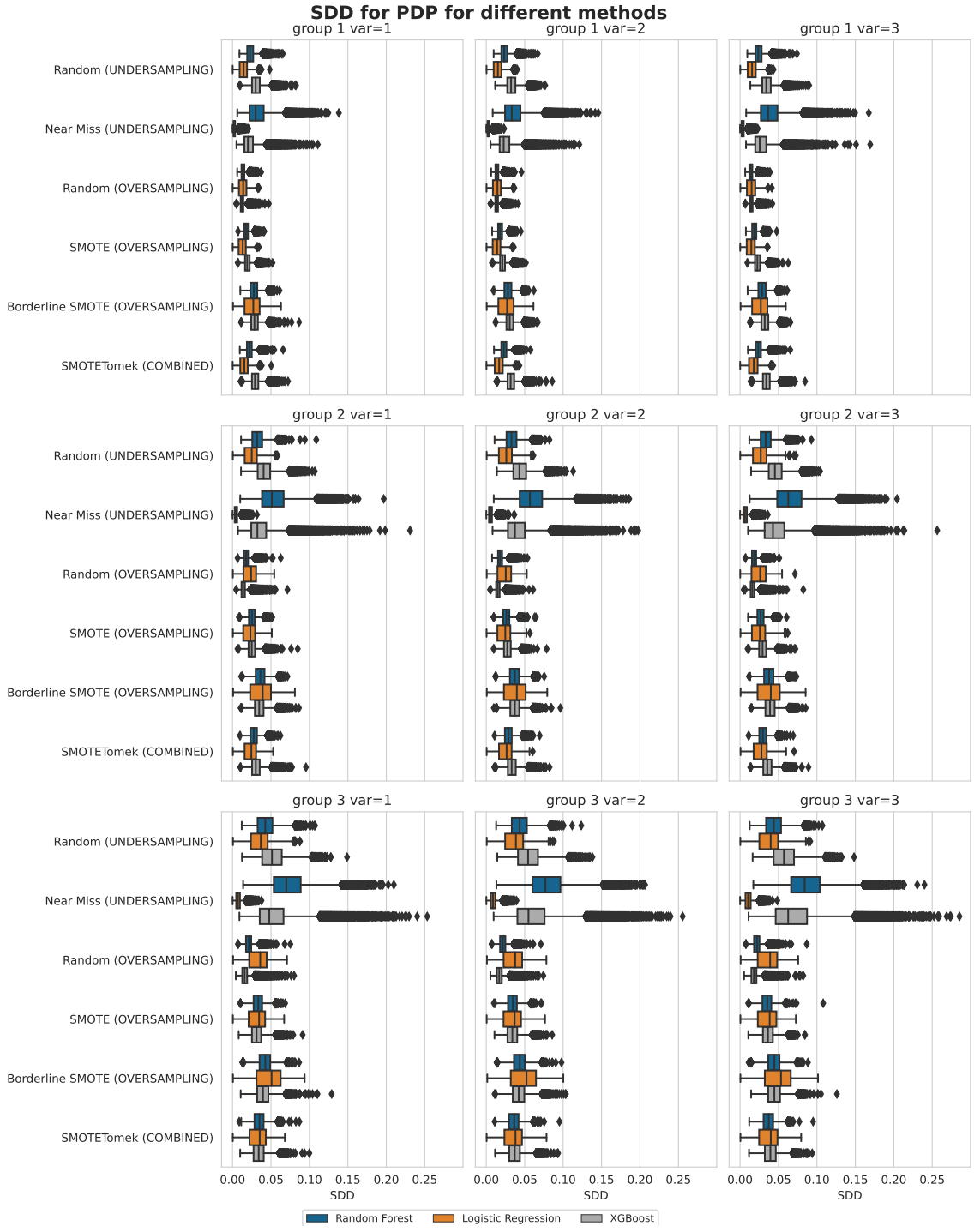


Figure 2: SDD results based on partial dependence profiles of the models trained on the simulated dataset. `group` represents the β_0 values (group i: $\beta_{0i} = [1.5, 2.5, 3.5, 4.5]$), and `var` represents the variance of the error term ($var_j = [1, 2, 3]$)

from the chart that balancing does improve model performance in terms of the evaluation metric. This confirms the results of many studies showing that these methods have a positive influence on the predictive power of the models. The biggest change observed is in the case of XGBoost, which has evolved from the worst to the best model after balancing. Nevertheless, there are some cases where the predictive power of the model decreased significantly. These outliers are the effect of the Near Miss method (Figure 3(b)). It is the only one that in some cases prevents the models from learning from the data. Apart from that, Figure 3(b) suggests that the best results were obtained after applying the **Random Undersampling** technique and that the Random Forest model benefited from it the most.

The SDD values for PDP and ALE profiles between models trained on the original and the balanced datasets are presented in Figure 4. Firstly, the SDD values for PDP and ALE are very alike in these plots. This means that either approach can be used to compare models. Secondly, all balancing methods cause significant changes in the behavior of the Logistic Regression models. Consequently, it can be concluded that this model is biased toward a balanced distribution. Moreover, the Near Miss method, which was the worst in terms of balanced accuracy, also causes the largest changes in SDD values. On the other hand, the XGBoost models have changed the least after applying Random Oversampling. This means that it is safe to use this technique when training an XGBoost model.

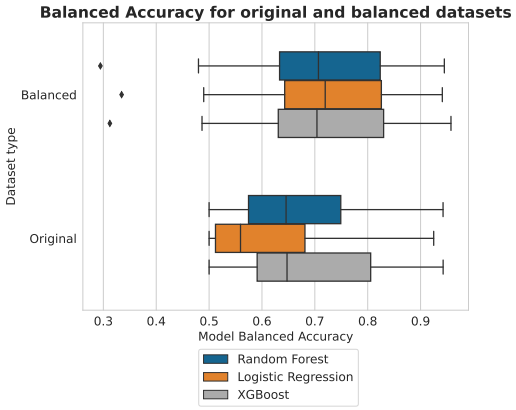
An example of how model behavior can change after balancing is illustrated in Figure 6. It presents ALE profiles for Random Forest models trained on the *wilt* dataset. It can be seen that the profiles for Random Undersampling and Near Miss methods have completely different characteristics compared to the original line.

The results of the Wilcoxon test, which compares the Variable Importance profiles of the models trained on the original and balanced datasets, are visualized in Figure 5. This plot confirms earlier observations that the **Near Miss** method causes the most changes in the model behavior. On the other hand, the XGBoost algorithm seems to be the most robust, as it has the largest fraction of accepted tests in all cases. However, it should be noted that the results obtained by the VI and PDP/ALE methods are not coherent. For example, the smallest SDD values were present for the XGBoost algorithm after applying **Random Oversampling**. The chart in Figure 5 does not depict that - the larger fraction of the accepted tests has Random Forest.

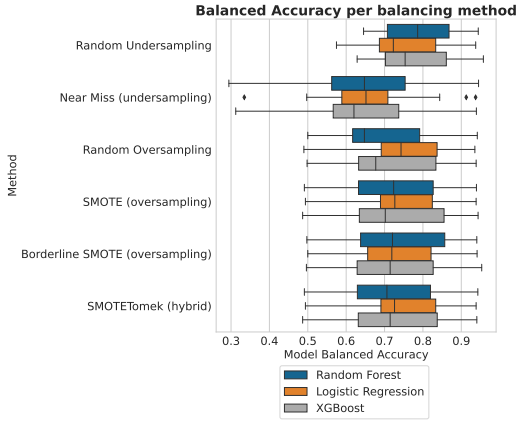
4.4. Performance gain plot

We proposed the performance gain plot that shows the relationship between the models and the balancing methods considered. The performance gain is given in the x-axis in terms of balanced accuracy and the ASDD values, which show the model behavior change based on the PDP or ALE profiles, are given in the y-axis. On such a scatterplot, we can compare two types of changes: in performance and behavior. The higher values on the x-axis and the lower the values on the y-axis is better. Conversely, high model behavior changes and low prediction performance gain. In this direction, we examined the changes in model behavior versus model performance gain by using the performance gain plot on simulated and real datasets.

In Figure 7, the most important observation in the case of Random Forest and XGBoost, the Near Miss method was the one with the highest behavior changes and the largest per-

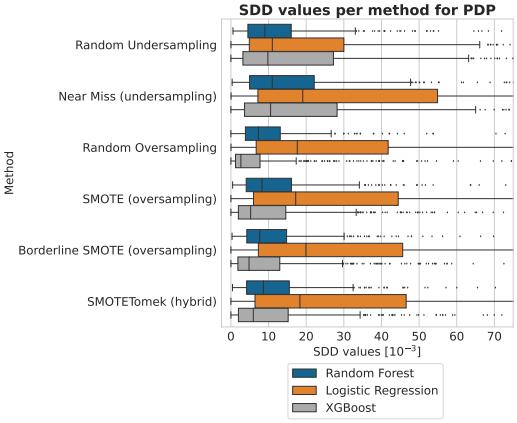


(*a*) Balanced Accuracy for different models on balanced and original datasets.

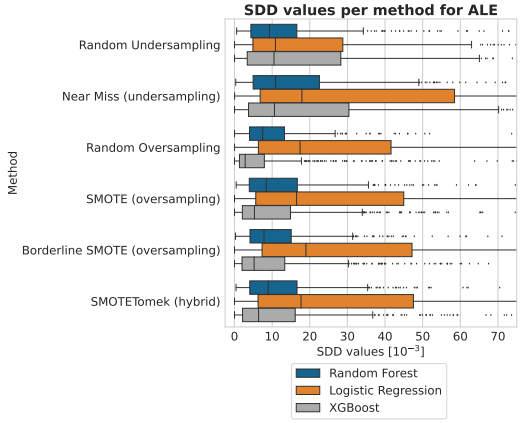


(*b*) Balanced Accuracy for different methods of balancing datasets.

Figure 3: Balanced Accuracy results of the real dataset experiments.



(*a*) The SDD values per balancing method for Partial Dependence Profiles.



(*b*) The SDD values per balancing method for Accumulated Local Effects Profiles.

Figure 4: The SDD results of the real dataset experiments.

formance change. On the contrary, the Random Undersampling technique had the highest performance gain, but still at the cost of behavior change. Another conclusion is that the ASDD values tend to be lower for Logistic Regression than for other models. Consequently, it can be concluded that the Near Miss method is the riskier method in terms of model behavior change for Random Forest and XGBoost models.

Figure 8 shows that the most important observation in the case of Random Forest and XGBoost is that for some datasets, the Near Miss method was the one with the highest behavior changes and the largest performance decrease. On the other hand, the Random Undersampling technique had the highest performance gain, but still at the cost of behavior

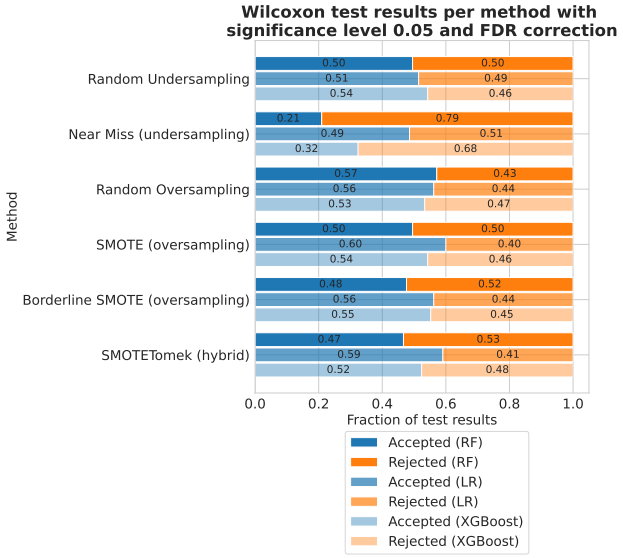


Figure 5: Wilcoxon test results with the significance level of 0.05 and FDR correction.

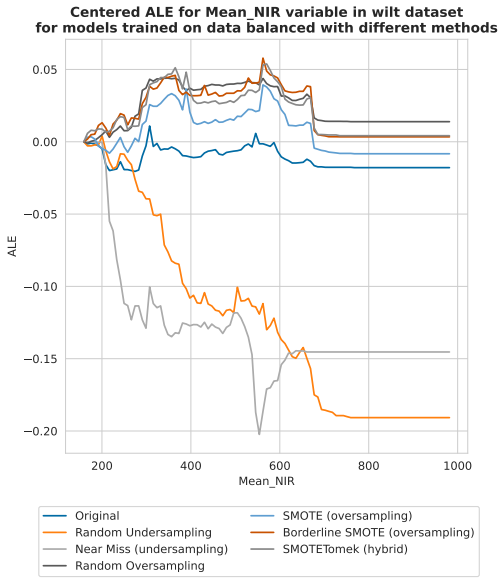


Figure 6: ALE plots for models trained on the wilt dataset balanced with different methods.

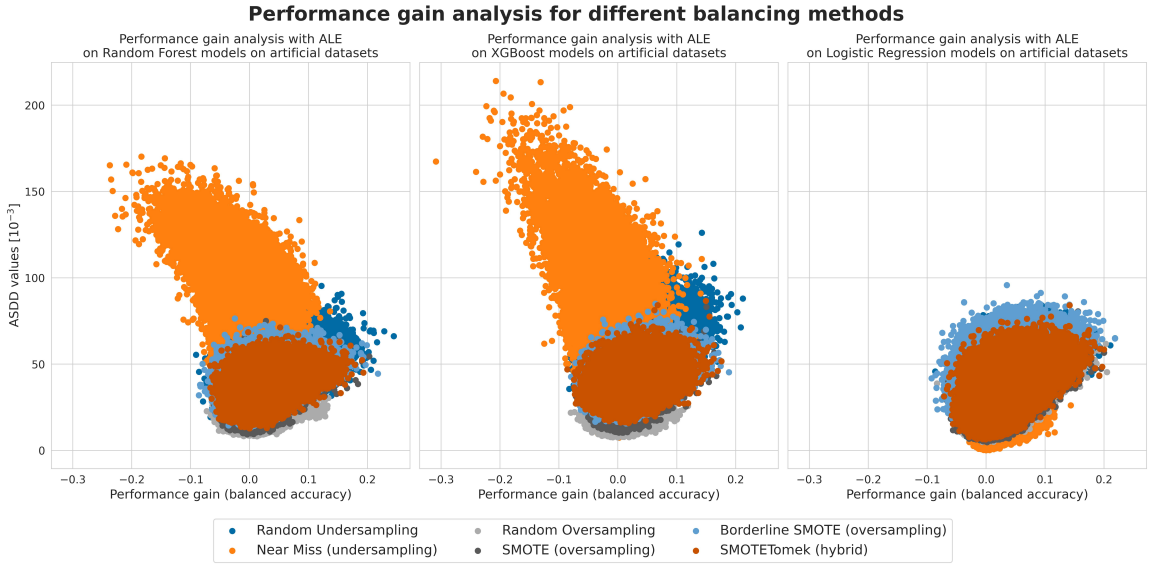


Figure 7: Performance gain plot for simulated datasets.

change. Another conclusion is that the ASDD values tend to be higher for Logistic Regression than for other models, while the decrease in the balanced accuracy value is almost not observed (or only to a small extent). Therefore, it can be concluded that data balancing

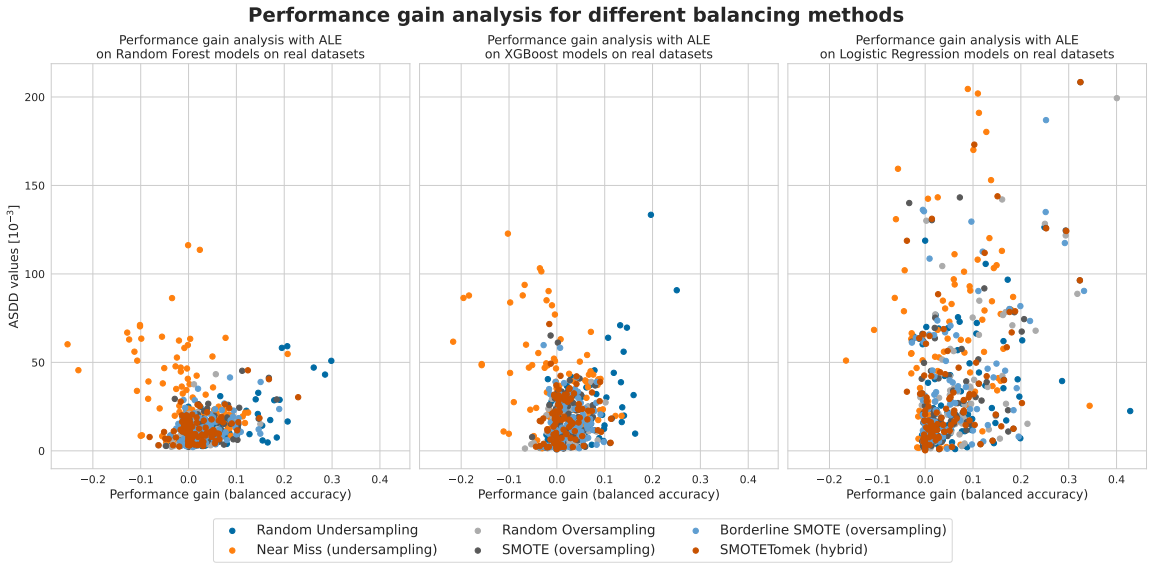


Figure 8: Performance gain plot for real datasets.

in the case of Logistic Regression is safer in terms of performance, but riskier in terms of behavior change.

5. Conclusions

In this paper, we investigated the impact of balancing methods on model behavior in imbalanced datasets. We conducted experiments on both simulated and real datasets to measure the impact of different balancing methods on model behavior and performance by using `edgaro` package. Our results show that **Random Undersampling** is the most effective method for improving model performance, followed by all **Oversampling** methods. However, the **Near Miss** method does not always lead to better performance, especially when the imbalance ratio is high. We also observed that the impact of the balancing methods on model behavior varies depending on the algorithm. These findings are consistent with the results presented in Moniz and Monteiro (2021) about the *No Free Lunch* concept (Schaffer, 1994) for imbalanced ML. Thus, we propose to use the performance gain plot to select the optimal balancing method in terms of performance gain and model behavior change. Additionally, we introduced a comprehensive model framework and followed a simulation design similar to previous studies to generate simulated datasets with controlled imbalances. The results of our experiments on these datasets demonstrate that the negative impact of balancing methods on model behavior increases with higher variance and imbalance ratios of model predictions.

In conclusion, our paper provides insights into the trade-offs between model performance and behavior when dealing with imbalanced datasets. Future research can explore alternative balancing methods, such as cost-sensitive learning, or combine multiple methods to further improve model performance and minimize changes in model behavior in imbalanced datasets.

Supplemental Materials

The materials for reproducing the experiments performed in Sec 4, the Python package, benchmark datasets, and the figures are accessible at a GitHub repository.

Acknowledgements

The work on this paper was carried out with the support of the Laboratory of Bioinformatics and Computational Genomics and the High-Performance Computing Center of the Faculty of Mathematics and Information Science, Warsaw University of Technology under computational grant number A-22-09. Also, it is financially supported by the NCN Sonata Bis-9 grant 2019/34/E/ST6/00052, and Eskisehir Technical University Scientific Research Projects Commission under grant no. 22ADP367.

References

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022.

Ismail Alarab and Simant Prakoonwit. Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques. *Data Science and Management*, 5(2):66–76, 2022.

Shideh Shams Amiri, Rosina O Weber, Prateek Goel, Owen Brooks, Archer Gandley, Brian Kitchell, and Aaron Zehm. Data representing ground-truth explanations to evaluate xai methods. *arXiv preprint arXiv:2011.09892*, 2020.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press, 2021.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. *arXiv preprint arXiv:1708.03731*, 2017.

Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 655–670. Springer, 2019.

Mustafa Cavus and Przemysław Biecek. Explainable expected goal models for performance analysis in football analytics. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9, 2022. doi: 10.1109/DSAA54385.2022.10032440.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Zejin Ding. *Diversified ensemble classifiers for highly imbalanced data learning and its application in bioinformatics*. PhD thesis, Georgia State University, 2011.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Qinghua Gu, Jingni Tian, Xuexian Li, and Song Jiang. A novel random forest integrated model for imbalanced data classification problem. *Knowledge-Based Systems*, 250:109050, 2022.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

Yazhe Li, Niall Adams, and Tony Bellotti. A relabeling approach to handling the class imbalance problem for logistic regression. *Journal of Computational and Graphical Statistics*, 31(1):241–253, 2022.

Nuno Moniz and Vitor Cerqueira. Automated imbalanced classification via meta-learning. *Expert Systems with Applications*, 178:115011, 2021.

Nuno Moniz and Hugo Monteiro. No free lunch in imbalanced learning. *Knowledge-Based Systems*, 227:107222, 2021.

Carlos Ortega Vázquez, Jochen De Weerdt, et al. Hellinger distance decision trees for pu learning in imbalanced data sets. *Machine Learning*, pages 1–32, 2023.

Aum Patil, Aman Framewala, and Faruk Kazi. Explainability of smote based oversampling for imbalanced dataset problems. In *2020 3rd international conference on information and computer technologies (ICICT)*, pages 41–45. IEEE, 2020.

Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 444–448. IEEE, 2022.

Mirka Saarela and Susanne Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3:1–12, 2021.

Cullen Schaffer. A conservation law for generalization performance. In *Machine Learning Proceedings 1994*, pages 259–265. Elsevier, 1994.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

Prabhant Singh and Joaquin Vanschoren. Automated imbalanced learning. *arXiv preprint arXiv:2211.00376*, 2022.

Adrian Stańdo. Impact of data balancing on model behavior with explainable artificial intelligence tools in imbalanced classification problems, 2023. URL https://github.com/MI2DataLab/MI2-prace-dyplomowe/blob/master/Prace/2023_Adrian_Sta%C5%84do.pdf.

Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022. doi: 10.1080/01605682.2020.1865846.

Qiyuan Zhang, Mark Hall, Mark Johansen, Vedran Galetic, Jacques Grange, Santiago Quintana-Amate, Alistair Nottle, Dylan M Jones, and Phillip L Morgan. Towards an integrated evaluation framework for xai: an experimental study. *Procedia Computer Science*, 207:3884–3893, 2022.