# An Experimental Study on the Rashomon Effect of Balancing Methods in Imbalanced Classification

Mustafa Cavus[1] and Przemysław Biecek[2,3]

[1] Eskisehir Technical University, Department of Statistics, Turkiye
mustafacavus@eskisehir.edu.tr
[2] Warsaw University of Technology, Faculty of Mathematics and Information Science, Poland
[3] University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Poland

**Abstract.** Predictive models may generate biased predictions when classifying imbalanced datasets. This happens when the model favors the majority class, leading to low performance in accurately predicting the minority class. To address this issue, balancing or resampling methods are critical pre-processing steps in the modeling process. However, there have been debates and questioning of the functionality of these methods in recent years. In particular, many candidate models may exhibit very similar predictive performance, which is called the Rashomon effect, in model selection. Selecting one of them without considering predictive multiplicity which is the case of yielding conflicting models' predictions for any sample may lead to a loss of using another model. In this study, in addition to the existing debates, the impact of balancing methods on predictive multiplicity is examined through the Rashomon effect. It is important because the blind model selection is risky from a set of approximately equally accurate models. This may lead to serious problems in model selection, validation, and explanation. To tackle this matter, we conducted real dataset experiments to observe the impact of balancing methods on predictive multiplicity through the Rashomon effect. Our findings showed that balancing methods inflate the predictive multiplicity, and they yield varying results. To monitor the trade-off between performance and predictive multiplicity for conducting the modeling process responsibly, we proposed using the extended performance-gain plot for the Rashomon effect.

**Keywords:** Predictive multiplicity · Explainable AI · Model behavior · Model selection.

## 1 Introduction

One of the most common challenges in classification tasks is imbalancedness. In this case, models tend to produce biased predictions toward the majority class, resulting in significantly lower performance for the minority class [1]. This issue can be solved using methods that balance the majority and minority classes.

Many papers have been published to implement data balancing techniques for solving imbalance classification problems. A systematic study of these techniques can be found in Vargas et al. [2]. While the literature offers many balancing methods, none is universally superior [3]. Each method has advantages and disadvantages as oversampling-based methods may result in the model learning excessively, potentially leading to overfitting, conversely, undersampling-based methods may result in information loss. Thus, the most suitable one should be chosen regarding the task and the dataset's characteristics.

The bias can be boosted throughout the machine learning pipeline depending on how exactly data is preprocessed, in the same manner, the pre-processing of the data also raises the change in the model such as changes in the model parameters [4], the correlations between variables [5], the importance of variables [6], and the model fairness [7]. As a component of data pre-processing, balancing methods may affect the model behavior and therefore model selection. The strategy followed can lead to changes in the models, thereby diversifying the resulting set of models [8]. Despite various perspectives addressing these methods related to performance, only a few studies have delved into their impact on model behavior. Patil et al. [5] showed that the SMOTE method effectively addressed imbalance issues while maintaining the original correlations between variables in the model. In contrast, Alarab and Prakoonwit [6] found that these methods altered feature importance in experiments conducted on two real datasets using various SMOTE variants. Stando et al. [4] conducted experiments on 21 datasets to investigate the impact of balancing methods on model behavior, and they showed the significant effects of balancing methods through partial dependence plots. They suggested that while utilizing balancing methods, it is essential to consider both performance gain and model behavior change. Kamalov et al. [9] discussed that full balancing is unnecessary to achieve optimal results and proposed partial resampling as the imbalanced ratio of 1.25.

However, no studies have been conducted on the Rashomon effect of balancing methods on the model behavior. The Rashomon effect is a phenomenon wherein various models can model a dataset approximately equally accurately [10]. Although this provides many alternatives in model selection, these models may yield conflicting predictions for any sample, which results in predictive multiplicity [11]. Thus, the potential negative consequences of randomly selecting a model must be considered as a part of responsible machine learning, as it masks the understanding of model differences. It is important because the performance of a model is not enough, and we need to explore model behavior in the phase of model selection [12]. Hence, it is imperative to carefully consider the implications of the Rashomon effect in the model selection process. Since the important effect of pre-processing steps in model selection, we take into account the effect of balancing methods on model behavior is examined in the Rashomon set. To the best of our knowledge, this is the first paper related to the Rashomon effect of balancing methods on model behavior in terms of multiplicity.

To investigate the effect of balancing methods on model behavior, we focus on the following research questions: (RQ1) How do the balancing methods affect the predictive multiplicity of the models in the Rashomon set?, (RQ2) How do the balancing methods affect the variable importance order discrepancy of the models in the Rashomon set? (RQ3) Can the partial resampling be a solution against the model behavior change? and (RQ4) Can the extended version of the performance gain plot be a solution to monitor the trade-off between performance gain and multiplicity? In the remainder of this paper, we first provide some preliminaries about the Rashomon effect and predictive multiplicity in Sec. 2, then we present the experiments in Sec. 3 and discuss the results in Sec. 4 and conclusions in the last section.

## 2 Preliminaries

Let $X = [x_1, x_2, \ldots, x_p]$ be a data matrix of $n$ observations from $p$ variables, and $y$ be the response vector. Also, let $F = \{f \mid f\colon X \to y\}$ be the space of all predictive models which is called **Hypothesis Space**. Furthermore, let $L\colon F \to \mathbb{R}$ denote the loss function of the model $f$. The goal is to find $f \in F$ that minimizes the expected value of the loss function $L$:

$$f = \underset{f \in F}{\operatorname{argmin}}\, \mathbb{E}[L(f)] \tag{1}$$

Note that expected loss $\mathbb{E}[L]$ is approximated with empirical loss calculated on data $X$.

**Reference Model.** Let $\hat{F} \subset F$, the model with the minimum loss function that we have found from $\hat{F}$ is called the reference model, and we shall denote it as $f_R$. In other words

$$f_R = \underset{f \in \hat{F}}{\operatorname{argmin}}\, \mathbb{E}[L(f)]. \tag{2}$$

**Rashomon Set.** In a learning problem, for a given loss function $L$, a reference model $f_R$, and the **Rashomon parameter** $\varepsilon > 0$ which limits the set, the Rashomon set $R_{L,\varepsilon}(f_R)$ is defined as

$$R_{L,\varepsilon}(f_R) = \{f \in F \mid \mathbb{E}[L(f)] \leq \mathbb{E}[L(f_R)] + \varepsilon\}. \tag{3}$$

It is not possible to access all possible models in $F$, so we are interested in an empirical Rashomon set, which we will refer to as just Rashomon Set

$$\hat{R}_{L,\varepsilon}(f_R) = \{f \in \hat{F} \mid \mathbb{E}[L(f)] \leq \mathbb{E}[L(\hat{f}_0)] + \varepsilon\}. \tag{4}$$

The number of models in the Rashomon set is defined as the **Rashomon Set Size**. It can be used as a metric to measure the severity of the Rashomon effect [13]. A high value of this metric indicates the presence of many candidate models for the same task, thus a high Rashomon effect, while a low value suggests a small number of candidate models and a low Rashomon effect. However, this criterion is only calculated based on the number of models in the Rashomon set, making it superficial, and does not reveal the outcome of this effect at the observation or model level. Marx et al. [11] have addressed this by introducing the *ambiguity* and *discrepancy* metrics.

**Ambiguity.** The ambiguity of a Rashomon Set $R_{L,\varepsilon}(f_R)$ is the ratio of observations in the data $X$ that has conflicting predictions with that reference model $f_R$ and the other models in $R_{L,\varepsilon}(f_R)$

$$\alpha_\varepsilon(f_R) = \frac{1}{n}\sum_{i=1}^{n} \max_{f \in R_{L,\varepsilon}(f_R)} \mathbb{1}(f(x_i) \neq f_R(x_i)). \tag{5}$$

**Discrepancy.** The discrepancy of a Rashomon Set $R_{L,\varepsilon}(f_R)$ is the maximum ratio of conflicting observations between the reference model $f_R$ and the other models in $R_{L,\varepsilon}(f_R)$

$$\delta_\varepsilon(f_R) = \max_{f \in R_{L,\varepsilon}(f_R)} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f(x_i) \neq f_R(x_i)). \tag{6}$$

Ambiguity indicates the number of observations assigned conflicting predictions by any competing model in the Rashomon set. In contrast, discrepancy indicates the maximum number of predictions that could change if the model is replaced with a competing model. Both take values between zero and one. Fig 1 illustrates the computation of ambiguity and discrepancy. The computation of these metrics depends on access to models in the Rashomon set, which is computationally infeasible [14,15]. Using specific model classes and retraining models with different hyperparameter setups are the strategies to create the Rashomon set. In this paper, we use a new strategy based on AutoML because of the easy access to the approximately equally accurate models from an intended model class, aka hypothesis space.

## 3 Experiments

In this section, we conduct experiments to observe the Rashomon effect of balancing methods on model behavior in terms of multiplicity. We considered four balancing methods: *random oversampling*, *SMOTE* [16], *random undersampling*, and *near miss* [17]. The first two of these methods are based on the idea of oversampling and the rest is based on undersampling of the minority class. We used the `forester`, which is a tree-based AutoML tool [18] to create a Rashomon set. We prefer to use this approach because it is more useful than the other ways to access the hypothesis space. It also enables to control of the Rashomon set size using the parameters provided by the Bayesian optimization part of the tool. The Rashomon set is created on the tasks within the imbalanced benchmark dataset proposed by [4], as provided in Table 1. It is one of the imbalanced benchmark datasets consists several data from the various domains. The imbalance ratio (*#samples in majority class/#samples in minority class*) of the

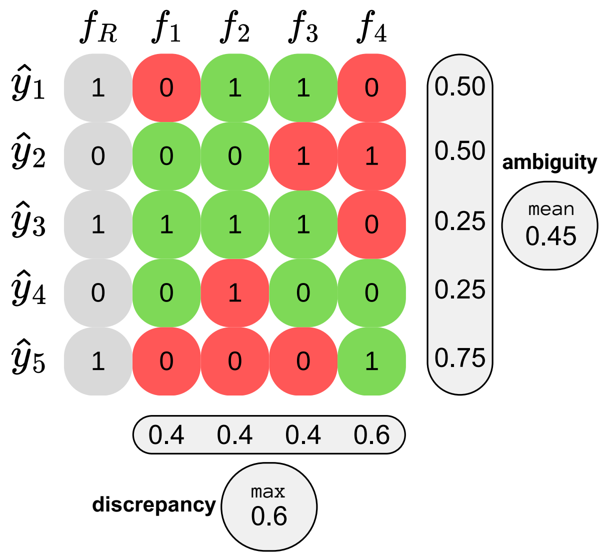


Fig. 1: **The illustration of calculations of *ambiguity* and *discrepancy*.** Assume that there are five models in the Rashomon set. We perform it on only five samples to simplify the illustration, but we would like to remind you that these two metrics are calculated on all samples. First column represents the reference model predictions $\hat{y}_i = f_R(X_i)$ for observations $i = 1, 2, 3, 4, 5$. The following columns show the models' predictions in the Rashomon set which are $f_1, f_2, f_3$, and $f_4$. The red cells indicate a conflict prediction and the green cells indicate a consistent prediction of the models in the Rashomon set. The maximum value of the ratio of the conflict predictions in the models called *discrepancy* and the mean value of the conflict predictions for observations called *ambiguity*.

datasets varies between 1.54 and 129.53. The Rashomon parameter $\varepsilon$ is taken as 0.05, and the resampling ratio (i.e., imbalanced ratio after balancing) varies as $\{1, 1.05, 1.10, 1.15, 1.20, 1.25\}$. Moreover, the number of optimization rounds `bayes_iter` is taken as 5, and the number of trained models `random_evals` is taken as 10 in the `train` function of the `forester`.

In the second part of the experiments, which focused on measuring the predictive multiplicity and investigating model behavior change, we utilized permutational variable importance. It has been frequently used an explainable artificial intelligence tool as a proxy in explanatory model analysis [19]. For example, Greenwell et al. [20] and Kozak and Biecek [21] have proposed methods for determining variable importance through the smoothness of these profiles. Moreover, Stando et al. [4], Zhang et al. [22], and Kobylinska et al. [23] have developed statistics to measure model behavior change calculating differences between profiles. In this study, we proposed a new metric which is called *Variable Importance Order Discrepancy* to measure dissimilarity between variable importance orders

Table 1: Imbalanced benchmark tabular datasets

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

of the models in the Rashomon set based on Kendall's $\tau$ correlation coefficient [24]:

$$\zeta_{\varepsilon}(f_R) = \max_{f \in R_{L,\varepsilon}(f_R)} \tau(f_R, f). \tag{7}$$

The variable importance order discrepancy of a Rashomon set measures the maximum dissimilarity of variable importance orders between the reference model and the other models in the Rashomon set. It provides the maximum dissimilarity of variable importance orders that could change if the model is replaced with a competing model. We use it to measure the model behavior change in terms of permutational variable importance for conducting the model selection responsibly. In addition, we used statistical plots [25], and statistical tests Kruskal-Wallis [26], Friedman [27], and Dunnett's pairwise test [28] to summarize and statistically evaluate the findings. The results are given in the following section.

## 4 Results

In this section, the experiments are conducted to explore the research questions considered in the paper.

**RQ1. How do the balancing methods affect the predictive multiplicity of the models in the Rashomon set?**

Monitoring the impact of balancing methods on the Rashomon set is crucial as multiplicity outlines the potential damage at the individual level induced by the arbitrary model selection [29]. Thus, we measure the multiplicity of the Rashomon sets after balancing the datasets.

The calculated ambiguity and discrepancy values of the Rashomon sets are given in Fig 2. The Rashomon zone visualizes Rashomon metrics in two dimensions for each Rashomon set created on original (imbalanced) and balanced datasets under varying resampling ratios. In this plot, if the zone created for the Rashomon set trained on the original dataset and the zone calculated for the Rashomon set trained on the balanced dataset by any balancing method completely overlap, we can visually interpret that the relevant balancing method does not change the Rashomon metrics, if not, it does. Although the Rashomon regions do not overlap exactly, they are not positioned very differently from each other on the discrepancy axis. However, the Rashomon regions of the balancing methods on the ambiguity axis do not overlap with the Rashomon regions of the original dataset. Balancing methods increase the ambiguity of Rashomon sets.

In addition to visual examinations of the plot, the effects of balancing methods on Rashomon metrics were compared using the Kruskal-Wallis test, and the results are given in Fig 3. The results show that there are statistical differences between the Rashomon metrics of the datasets because the p-value of the tests is lower than 0.05 for both ambiguity and discrepancy. The results of the pairwise test show the source of the difference is the original dataset. The Rashomon metrics of the original dataset are lower than the balanced datasets. It is concluded that the balancing methods increase the discrepancy and ambiguity of the Rashomon sets.

**RQ2. How do the balancing methods affect the variable importance order discrepancy of the models in the Rashomon set?**

The models in the Rashomon set may not use similar variables [30], and also the small changes in training data can produce large changes in the outputs [31]. Thus, it is important to check whether the balancing methods inflate multiplicity and we investigate the effect of balancing methods on model behavior. The variable importance order is an important way to measure the model behavior change. Here, we used the variable importance order discrepancy to measure the discrepancy of the variable importance orders in the Rashomon set. Thus, we want to carry out the model selection process responsibly, being aware of how a model selected from the Rashomon set behaves differently from other models in terms of variable importance order.

The effects of balancing methods on variable importance order discrepancy are compared using the Kruskal-Wallis test and the results are given in Fig 4. There is no statistically significant difference between the median of the variable importance order discrepancy values of the Rashomon set because the p-value of the Kruskal-Wallis test is not lower than 0.05.

**RQ3. Can the partial resampling be a solution against the model behavior change?**

The partial resampling is proposed to mitigate the bias of balancing methods. Here, we use partial resampling to check whether it can be a solution to mitigate the effect of balancing methods on the model behavior.

The distribution of the Rashomon metrics *ambiguity*, *discrepancy*, and *variable importance order discrepancy* is given for balanced and partially balanced datasets in Fig 5. No pattern between the metrics over resampling ratios is seen. We conduct the Friedman test to test the effect of resampling ratios on the Rashomon metrics. The comparison of the Rashomon metrics over the balancing methods under the block effect of resampling ratios was performed individually using Friedman's test showing statistically significant differences, and the results are as follows: $\chi^2_{(5)} = 6.8286$, $p = 0.2337$, $\chi^2_{(5)} = 9$, $p = 0.1091$, and $\chi^2_{(5)} = 3.7429$, $p = 0.5870$, respectively. There is no statistically significant difference between the Rashomon metrics over resampling ratios because the p-values are not lower than 0.05. Thus, partial resampling is not a solution for the multiplicity problem of balancing methods.

**RQ4. Can performance gain plot be a method to monitor the trade-off between performance gain and multiplicity?**

We considered alternative solutions after determining that the partial resampling method was insufficient to combat the model behavior change. One of the potential solutions is the *performance-gain plot* which is proposed to monitor the trade-off between the effect of balancing methods and performance gain [4]. Here, we expand it for Rashomon metrics to monitor the multiplicity in the Rashomon set over performance gain. The performance-gain plot for Rashomon metrics is given in Fig 6. The oversampling-based resampling methods *Oversampling* and *SMOTE* improve the performance higher than others in all cases. One of the interesting findings is that although the performance gain increases when the data is made more balanced with the *Near miss* method, the ambiguity, and discrepancy also increase. However, it should be noted that the performance gain here is not noteworthy. Although the *Oversampling* and *SMOTE* methods seem to provide close performance gains, it can be said that the *Oversampling* method leads Rashomon sets consisting of models with similar variable importance order.

## 5 Conclusions

In this paper, the effect of balancing methods on predictive multiplicity was investigated by addressing different **RQs**. It is important because exploring the Rashomon set provides some advantages and challenges. It can be utilized to select models that satisfy additional properties without compromising accuracy, such as fairness, interpretability, and stability [32]. However, it may be problematic because the models in the Rashomon set rely on different variables and the blind selection of the model can produce more profitable or unfavorable predictions to the individuals [33].

We conduct experiments on benchmark datasets with varying imbalanced ratios. For the **RQ1**, we found that balancing methods increase Rashomon metrics. In the **RQ2**, we investigated the effect of balancing methods with the newly proposed metric we proposed on the model behavior through variable importance order and did not observe a statistically significant order change that contradicts the findings of Alarab and Prakoonwit [6]. However, considering that they conducted their experiments on only 2 data sets, we think the results we obtained on a larger number of datasets are more comprehensive and reliable. The results we obtained above that the balancing methods inflate the Rashomon metrics making it necessary to develop solutions. First, in the **RQ3**, we examined whether the partial resampling approach suggested by Kamalov et al. [9] could be a solution. However, we found that partial resampling does not work. Then, by expanding the performance-gain plot and monitoring the balance between performance and Rashomon metrics, we showed that choosing the method that best suits the data could be a solution. With its easy-to-use structure, the performance-gain plot is quite flexible and expandable across the intended Rashomon metrics.

In conclusion, we would like to draw attention again to the recent discussions of the use of balancing methods with an additional dimension. Our experiments show that balancing methods increase the predictive multiplicity and increase the risks involved in the process of model selection from the Rashomon set. Therefore, researchers should consider the risk of multiplicity when using balancing methods, which are a haven in imbalanced classification problems. The use of the extended performance gain plot for the Rashomon effect is important to monitor this issue and conduct the modeling process responsibly.

## 6 Further Research

There are still opportunities for further exploration in this field. To mitigate the predictive multiplicity, identifying the source of the Rashomon effect is a hot topic [33]. First, Komorniczak et al. [34] stated in their study that oversampling methods increase data complexity. The question of whether there is a connection between data complexity and the Rashomon effect can be investigated, e.g., "Does the increase in data complexity due to oversampling methods lead to an increase in the Rashomon effect?". Second, similar to the study by Junior and Pisani [35], it can be investigated whether cost-sensitive methods, such as resampling methods, have a lesser impact on the Rashomon effect. Third, special cases of imbalance situations such as class overlap, and small disjunct, as mentioned by authors such as Garcia et al. [36] and Prati et al. [37], can also be examined.

## Supplemental Materials

The materials for reproducing the experiments performed and the benchmark datasets are given in the repository: `https://github.com/mcavs/ECML2024_Imbalanced_Rashomon_Paper`.

## References

1. Khoshgoftaar, T. M., Seiffert, C., Van Hulse, J., Napolitano, A., Folleco, A.: Learning with limited minority class data, In Sixth International Conference on Machine Learning and Applications, pp.348–353 (2007).
2. Vargas, W., Aranda, J. A. S., dos Santos Costa, R., da Silva Pereira, P. R., Victória Barbosa, J. L.: Imbalanced data pre-processing techniques for machine learning: a systematic mapping study. Knowledge and Information Systems, **65**(1), pp. 31–57 (2023)
3. Moniz, N., Monteiro, H.: No free lunch in imbalanced learning. Knowledge-Based Systems **227**, 107222 (2021).
4. Stando, A., Cavus, M., Biecek, P.: The effect of balancing methods on model behavior in imbalanced classification problems. arXiv preprint arXiv:2307.00157 (2023)
5. Patil, A., Framewala, A., Kazi, F.: Explainability of SMOTE-based oversampling for imbalanced dataset problems. In 2020 3rd International Conference on Information and Computer Technologies, pp. 41–45 (2020)
6. Alarab, I., Prakoonwit, S.: Effect of data resampling on feature importance in imbalanced blockchain data - comparison studies of resampling techniques. Data Science and Management, **52**, pp. 66–76 (2022).
7. Simson, J., Pfisterer, F., Kern, C.: One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. arXiv preprint arXiv:2308.16681 (2024).
8. Watson-Daniels, J., Barocas, S., Hofman, J. M., Chouldechova, A.: Multi-target multiplicity: flexibility and fairness in target specification under resource constraints, In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 297–311 (2023).
9. Kamalov, F., Atiya, A. F., Elreedy, D.: Partial resampling of imbalanced data. arXiv preprint arXiv:2207.04631 (2022).
10. Breiman, L.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statistical science **16**(3), 199–231 (2001)
11. Marx, C., Calmon, F., Ustun, B.: Predictive multiplicity in classification. In International Conference on Machine Learning, pp. 6765–6774. PMLR (2020)
12. Biecek, P., Baniecki, H., Krzyznski, M., Cook, D.: Performance is not enough: a story of the Rashomon's quartet. arXiv preprint arXiv:2302.13356 (2023)
13. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, pp.1–85 (2022).
14. Hsu, H., Li, G., Hu, S.: Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. arXiv preprint arXiv:2402.00728 (2024).
15. Donnelly, J., Katta, S., Rudin, C., Browne, E.: The Rashomon Importance Distribution: Getting RID of Unstable, Single Model-based Variable Importance. Advances in Neural Information Processing Systems, 36 (2024).

16. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, pp.321–357 (2002).
17. Mani, I., Zhang, I.: kNN approach to unbalanced data distributions: a case study involving information extraction, In Proceedings of the workshop on learning from imbalanced datasets, 126, pp. 1–7 (2003).
18. Kozak, A., Ruczyński, H.: forester: A Novel Approach to Accessible and Interpretable AutoML for Tree-Based Modeling, AutoML Conference (2023).
19. Biecek, P., Burzykowski, T.: Explanatory Model Analysis. Chapman and Hall/CRC, New York (2021).
20. Greenwell, B. M., Boehmke, B. C., Gray, B.: Variable importance plots-an introduction to the vip package. R Journal, **21**(1), pp. 343–366 (2020).
21. Kozak, A., Biecek, P.: vivo: Variable Importance via Oscillations. R package version 0.2.1, https://CRAN.R-project.org/package=vivo, 2020.
22. Zhang, Y., Xu, F., Zou, J., Petrosian, O. L., Krinkin, K. V.: XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction. In the Second International Conference on Neural Networks and Neurotechnologies, pp. 13–16 (2021)
23. Kobylińska, K., Krzyziński, M., Machowicz, R., Adamek, M., Biecek, P.: Exploration of Rashomon set assists explanations for medical data. arXiv preprint arXiv:2308.11446 (2023)
24. Kendall, M.G.: A new measure of rank correlation. Biometrika, 30, pp.81–93 (1938).
25. Patil, I.: Visualizations with statistical details: The ggstatsplot approach, Journal of Open Source Software, 6(61), pp. 3167 (2021).
26. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, pp.583–621 (1952).
27. Kruskal, W.H., Wallis, W.A.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32, pp.675–701 (1937).
28. Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, pp.1096–1121 (1955).
29. Hsu, H., Calmon, F.: Rashomon capacity: A metric for predictive multiplicity in classification. Advances in Neural Information Processing Systems, 35, pp.28988–29000 (2022).
30. Poiret, C., Grigis, A., Thomas, J., Noulhiane, M.: Can we Agree? On the Rashomon Effect and the Reliability of Post-Hoc Explainable AI. arXiv preprint arXiv:2308.07247 (2023).
31. Oh, S., Ustun, B., McAuley, J., Kumar, S.: Rank list sensitivity of recommender systems to interaction perturbations. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 1584–1594 (2022).
32. Paes, L. M., Cruz, R., Calmon, F. P., Diaz, M.: On the inevitability of the Rashomon effect. In 2023 IEEE International Symposium on Information Theory, pp. 549–554 (2023).
33. Meyer, A.P., Albarghouthi, A., D'Antoni, L.: The dataset multiplicity problem: How unreliable data impacts predictions. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp.193–204 (2023).
34. Komorniczak, J., Ksieniewicz, P., Woźniak, M.: Data complexity and classification accuracy correlation in oversampling algorithms. In Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications, pp. 175–186, PMLR (2022).

35. Junior, J. D. S. F., Pisani, P. H.: Performance and model complexity on imbalanced datasets using resampling and cost-sensitive algorithms. In Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications, pp. 83—97, PMLR (2022).
36. Garcia, V., Sánchez, J. S., Mollineda, R. A.: On the effectiveness of pre-processing methods when dealing with different levels of class imbalance. Knowledge-Based Systems, 25(1), pp. 13–21 (2012).
37. Prati, R. C., Batista, G. E., Silva, D. F.: Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowledge and Information Systems, 45, pp. 247–270 (2015).

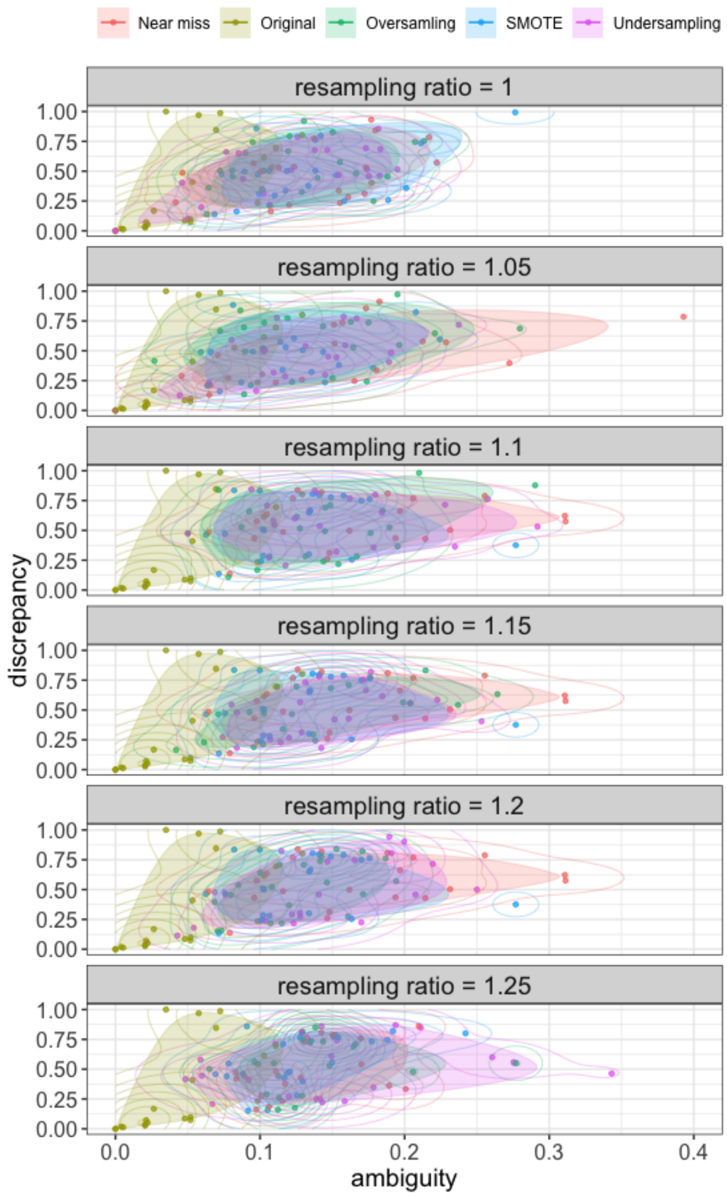


Fig. 2: **The zone plot of the Rashomon metrics *ambiguity* and *discrepancy* for different balancing methods and various resampling ratios.** The zones show two-dimensional regions where metrics' values are dense. The value of 1.25 means the frequency of the majority class over the frequency of the minority class. Zones being close to zero on both axes indicate low severity of multiplicity, and moving away from them indicates increasing severity of multiplicity.

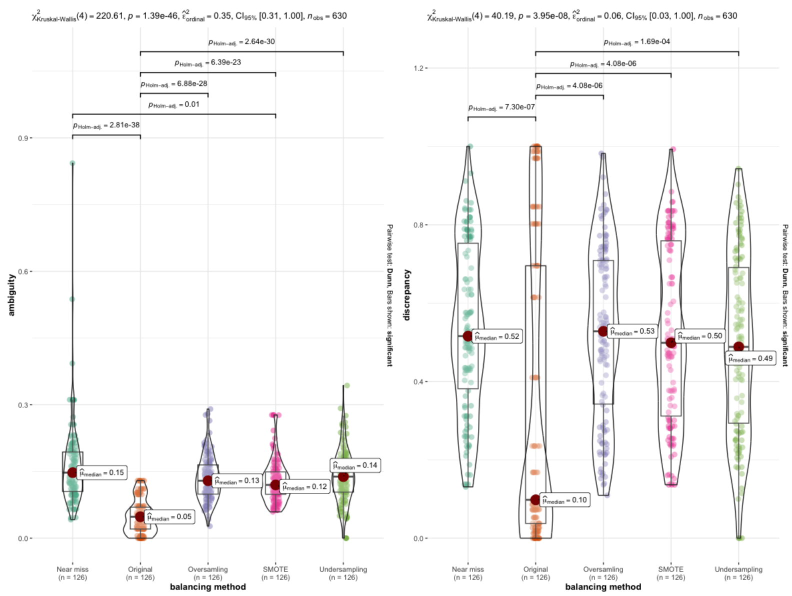


Fig. 3: **The distribution plots of the Rashomon metrics *ambiguity* and *discrepancy* for different balancing methods.** The plots consist of the results of the statistical tests Kruskal-Wallis and Dunn's Pairwise tests. The reference bars above each violin indicate statistically significant differences between the medians of groups and the corresponding statistical information.

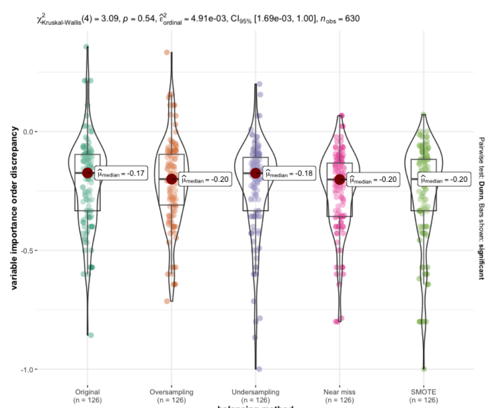


Fig. 4: **The distribution plots of the Rashomon metric *variable importance order discrepancy* for different balancing methods.** The plots consist of the results of the only Kruskal-Wallis test. Because there is no difference between groups, no need to conduct any pairwise comparison test.

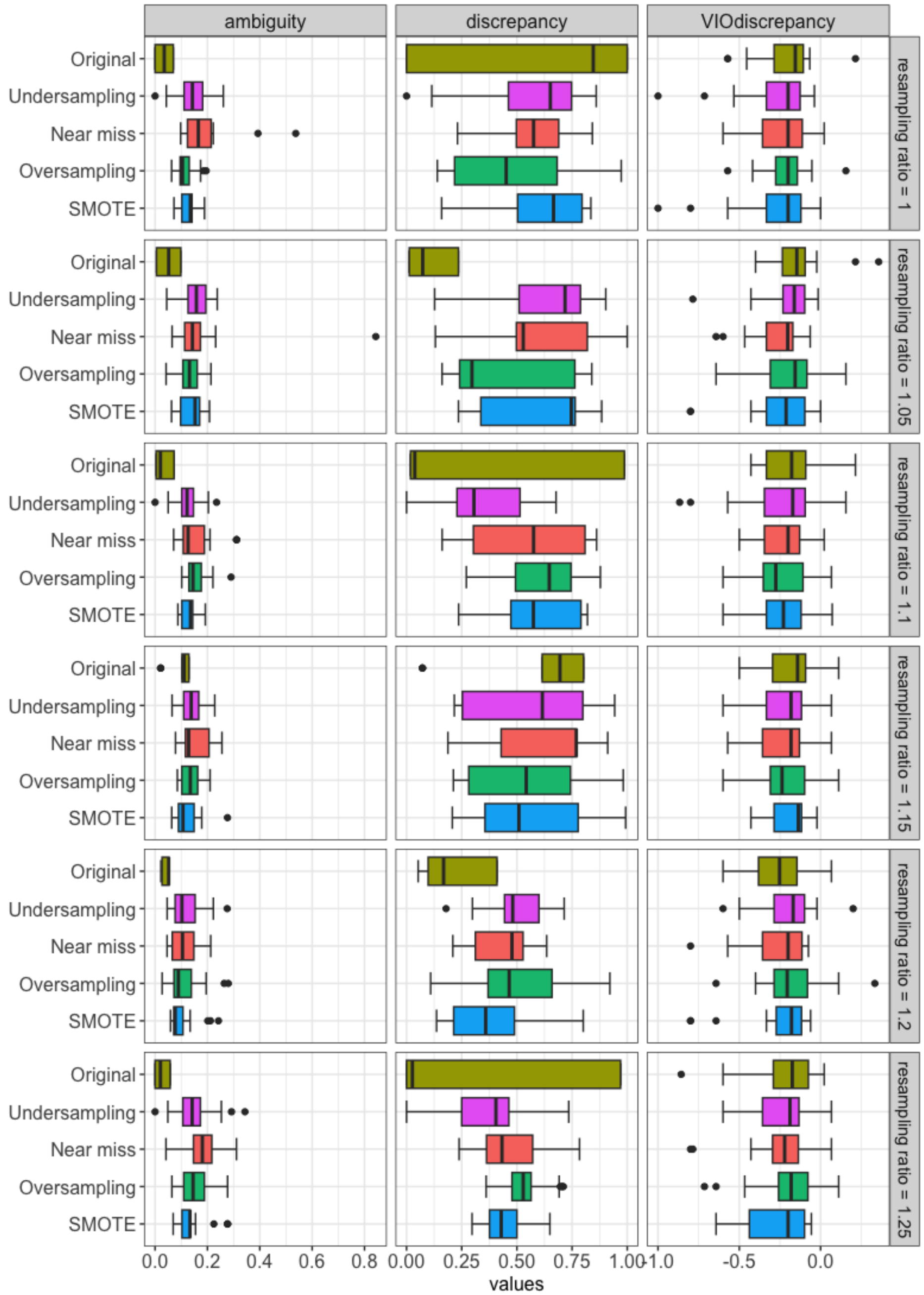


Fig. 5: **The distribution plots of the Rashomon metric *variable importance order discrepancy* for different balancing methods and varying partial resampling ratios.**

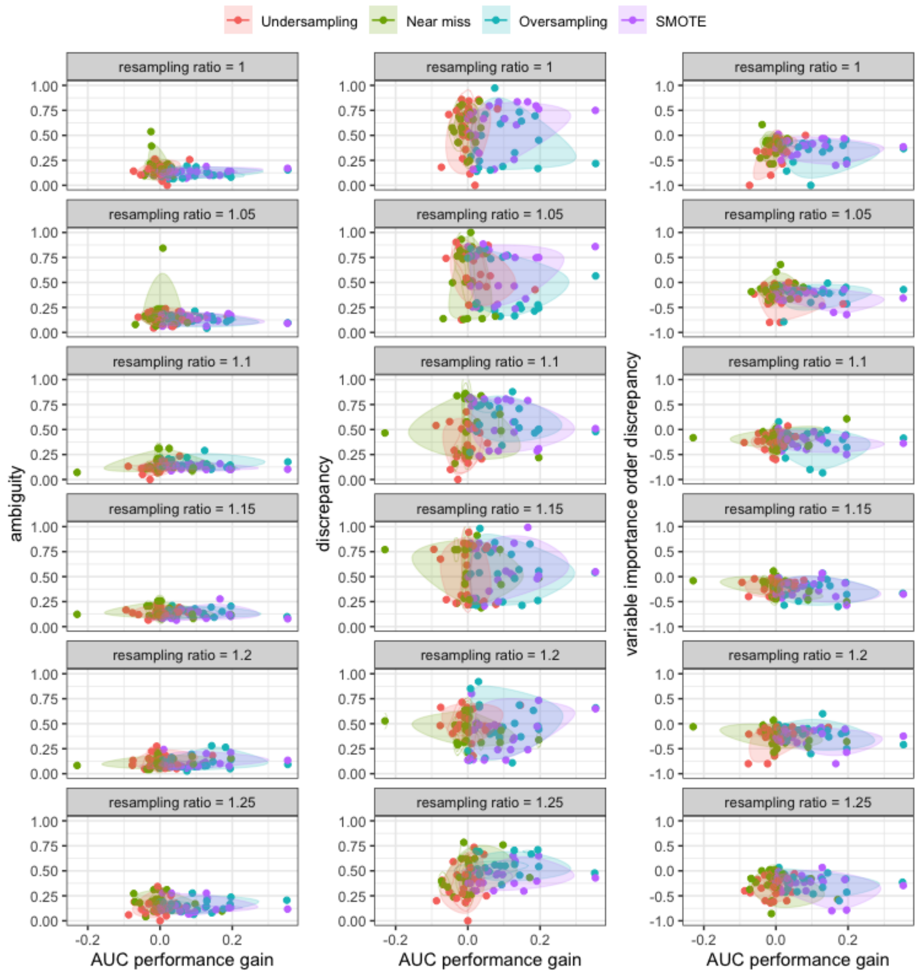


Fig. 6: **The performance gain plots of *ambiguity*, *discrepancy*, *variable importance order discrepancy* for different balancing methods and varying partial resampling ratios.** The horizontal axis shows the performance gain in terms of AUC. The zero indicates no gain, and the negative values indicate the performance loss. The vertical axes are limited between zero and one for *ambiguity* and *discrepacny*, but it is between minus one and one for *variable importance order discrepancy*. The moving of the zones towards the positive way on the horizontal axis indicates an increase in performance gain, and moving towards the negative way on the vertical axis indicates a decrease in the multiplicity.