Data poisoning on PDP with marginal distributions constraints

Julia Chylak JC394241@STUDENTS.MIMUW.EDU.PL
Aleksandra Mysiak AN.MYSIAK@STUDENT.UW.EDU.PL
Krzysztof Tomala KT439979@STUDENTS.MIMUW.EDU.PL
University of Warsaw, Poland

Abstract

In this project, we quantify the difference in impact of partial dependence-directed data poisoning on PD and ALE explanations. We compare this effect across multiple datasets. Additionally, we introduce a new, distribution-related loss term to the poisoning procedure. We show that it enables the procedure to significantly disturb the explanations while keeping the poisoned data samples' marginal distributions similar to the original ones.

1. Introduction

Partial dependence plots (PDPs) (Friedman, 2001) are a widely used method for understanding how a model is making predictions. Although widely used, PDPs are vulnerable to being misled by manipulated data. This research assesses the extent to which PDPs can be fooled by intentionally altering the input data while keeping the data distribution unchanged to conceal the attack. Additionally, this study compares the results obtained on PDPs with those from another explanation method, Accumulated Local Effects (ALE) (Apley and Zhu, 2016), to provide an understanding of the limitations in fooling either method.

2. Methodology

The methodology used in our experiments is based on the gradient method introduced in Baniecki et al. (2021). For each dataset, we first train a simple network with 2 fully connected layers. We then utilise an Adam optimizer (Kingma and Ba, 2014) to modify the data sample so that the PDP calculated on the new sample changes. Our goal here is to create a disruption that is significant from a real-life point of view. As such, we focus not on maximising distance between original and poisoned explanations, but on reversing trends visible in the explanations. Our target explanation in each experiment is a negation of the original PD profile, which serves as a quick-to-compute proxy for trend reversal. For this part, we use an L2 loss.

Additionally, to make the disruptions more impactful, we aim to create data samples comparable to the original ones. We introduce an additional weighted loss term that measures differences between marginal distributions of all features – we call it the *distribution*

distance loss. Our final loss is

$$L(X_{original}, X_{changed}) = L_2\left(PD(X_{original}), PD(X_{changed})\right) + \texttt{dist\_weight} * dist\_loss(X_{original}, X_{changed}),$$

where to compute *dist_loss* we sort the features of $X_{original}$ and $X_{changed}$ and compute MSE between them.

Algorithm Compute distribution loss

1: $x_1 = sort(X_{original}, axis = 0)$
2: $x_2 = sort(X_{changed}, axis = 0)$
3: return mean$((x_1 - x_2)^2)$

2.1 Metrics

We measure differences between two types of explanations: PDP and ALE, calculating metrics between explanations calculated on the original data and on the poisoned sample. The target explanation is ignored here, although it is indirectly taken into account via its relation to the original explanation. Our main metrics are:

- L2 metric – standard L2 metric, featured in the loss function.
- Spearman's ρ – a standard measure of rank correlation. A negative value suggests at least partial reversal of monotonicity.

The ρ metric is the focus of this investigation. It relates directly to the notion of trend reversal that was chosen here as a proxy for change in human interpretation of explanations. It is also useful when comparing explanation types, since the other metrics are scale-dependent and, as such, are hard to compare between PDP and ALE.

2.2 Datasets

We limit our experiments to classification problems only, so that all the results are comparable in scale. The datasets used are: **adult** (Andras Janosi et al., 2017), **bike** (Fanaee-T and Gama, 2013), **heart** (Dua and Graff, 2017), **titanic** (Harrell Jr. and Cason, 1912) and **xor** (inspired by code samples in Baniecki et al. (2021), our only artificial dataset). All datasets except for bike have a binary reponse variable; for bike, we have adapted the response variable to indicate whether a value for a given time is above average.

3. Experimental results

We have carried out a series of experiments for 6 variables (from 5 datasets), 3 network sizes and 6 values of `dist_weight`, with 10 repetitions per experiment. Results of those experiments have been gathered into a graphical form.

In Figure 1 we can see the effects of attacks for PDP and ALE across multiple datasets, with `dist_weight` $= 0$. We can see that for all datasets, the impact on PDP is significantly larger than on ALE. However, the ρ values for ALE are negative for all but one variables, indicating at least a partial trend reversal also in this case.

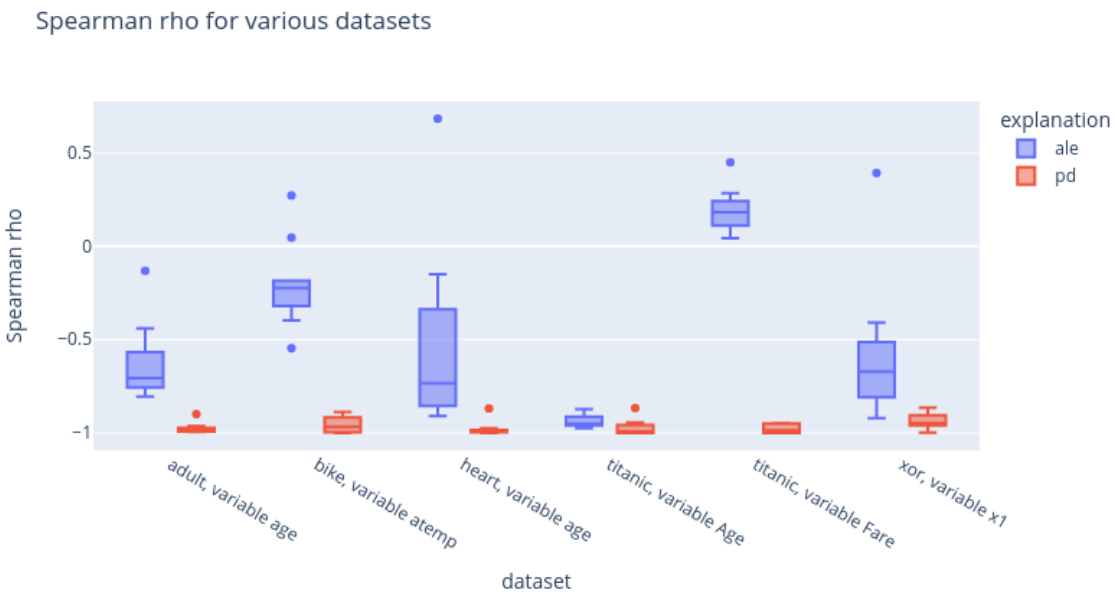


Figure 1: Boxplots of poisoning effects for different datasets, `dist_weight` $= 0$.

In Figure 2 we can see the results of a non-zero `dist_weight`. We can see that for all the features, the resulting distributions are much more realistic than without the additional loss, while trends in explanations remain reversed. More results regarding distribution distance can be found in the appendix.

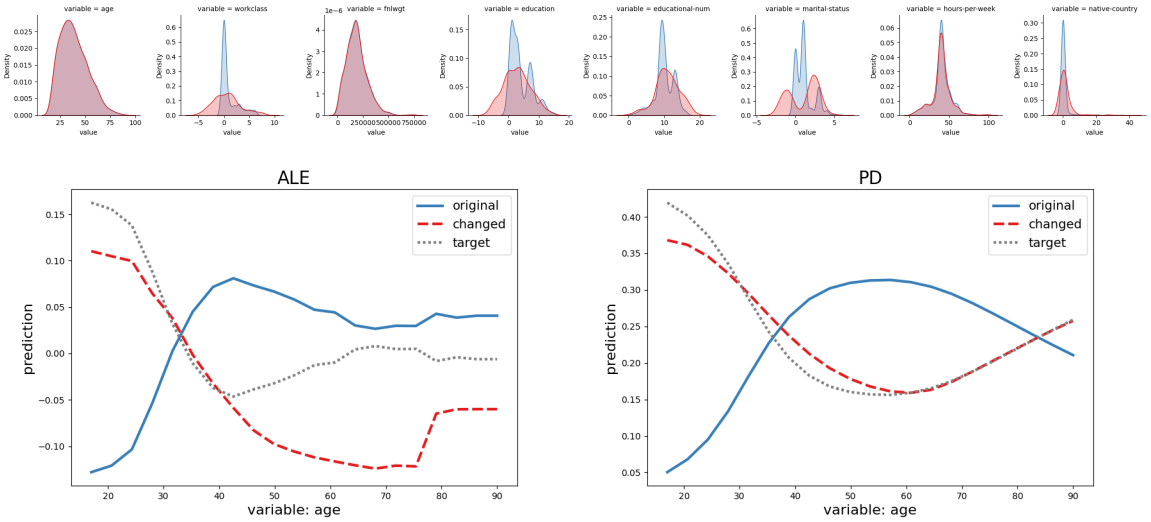
4. Conclusion

We have shown that regardless of the dataset used, both PD and ALE explanations are susceptible to PD-directed attacks – although the extent of the vulnerability may vary with dataset changes. As such, they should not be blindly trusted in scenarios when the data sample might have been manipulated, and even providing multiple explanation methods cannot guarantee validity of observable trends.

Additionally, we have also shown a method of limiting the impact of the poisoning procedure on features' marginal distributions. This indicates that simple, single-feature checks might not be enough to assert that a given data sample was not manipulated. Correlations between features were not considered here, and can be used as an indicator of data manipulation in our results.

Code for this work is available at `https://github.com/julkaztwittera/fooling-partial-dependence`.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.01

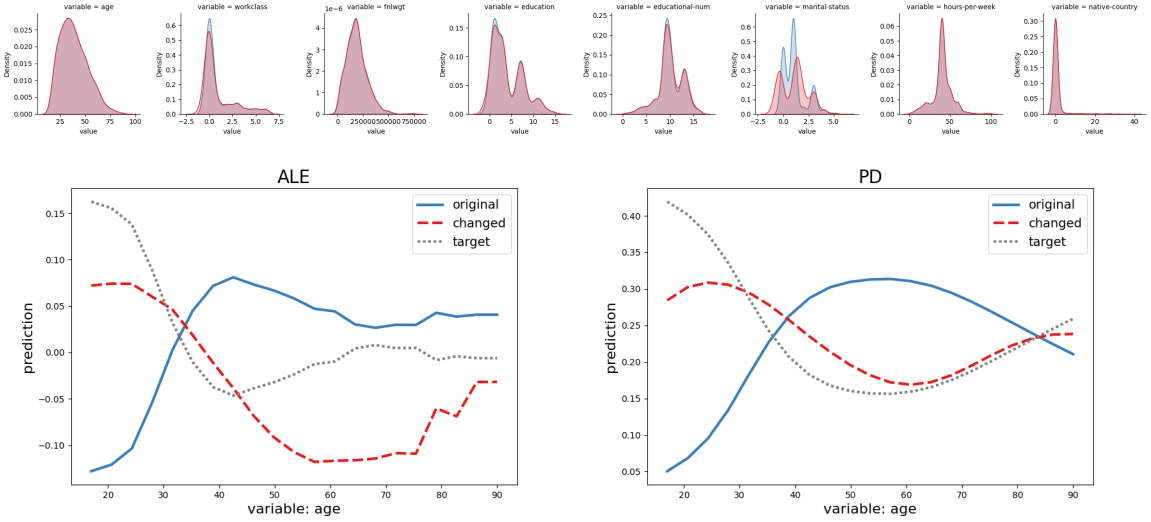


Figure 2: Comparison of poisoning results for the **adult** dataset, variable age.

References

M.D. Andras Janosi, M.D. William Steinbrunn, M.D. Matthias Pfisterer, and M.D. Ph.D. Robert Detrano. UCI Machine Learning Repository, Heart Disease Data Set, 2017. URL https://archive.ics.uci.edu/ml/datasets/adult.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models, 2016. URL https://arxiv.org/abs/1612.08468.

Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. Fooling partial dependence via data poisoning. *CoRR*, abs/2105.12837, 2021. URL https://arxiv.org/abs/2105.12837.

Dheeru Dua and Casey Graff. UCI machine learning repository, Adult Data Set, 2017. URL https://archive.ics.uci.edu/ml/datasets/heart+disease.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Frank E. Harrell Jr. and Thomas Cason. Titanic Dataset. 1912. URL https://www.kaggle.com/competitions/titanic/data.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

5. Additional results

In this appendix, we present additional results related to the distribution distance part of the loss function.

In Figure 3, we can see the impact of the value of `dist_weight` on attack results for different datasets, as measured by the ρ metric for PD plots. We can see that there is a trade-off between maintaining the marginal distributions and attack results. However, for all datasets, we can still select a non-zero weight that will not disturb results – our method provides an additional benefit without disturbing the results.

In Figures 4, 5, 6, 7 and 8 we can see comparisons of constrained and non-constrained attack results across all our datasets. We can see that in each dataset, there is at least one example where a trend reversal can be achieved with limited disturbance to the marginal distributions.

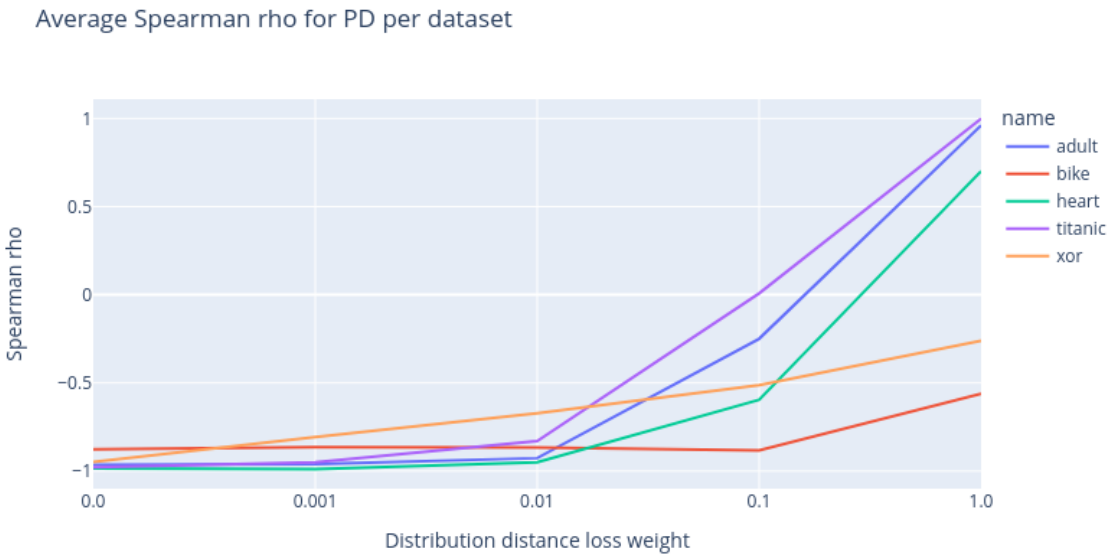
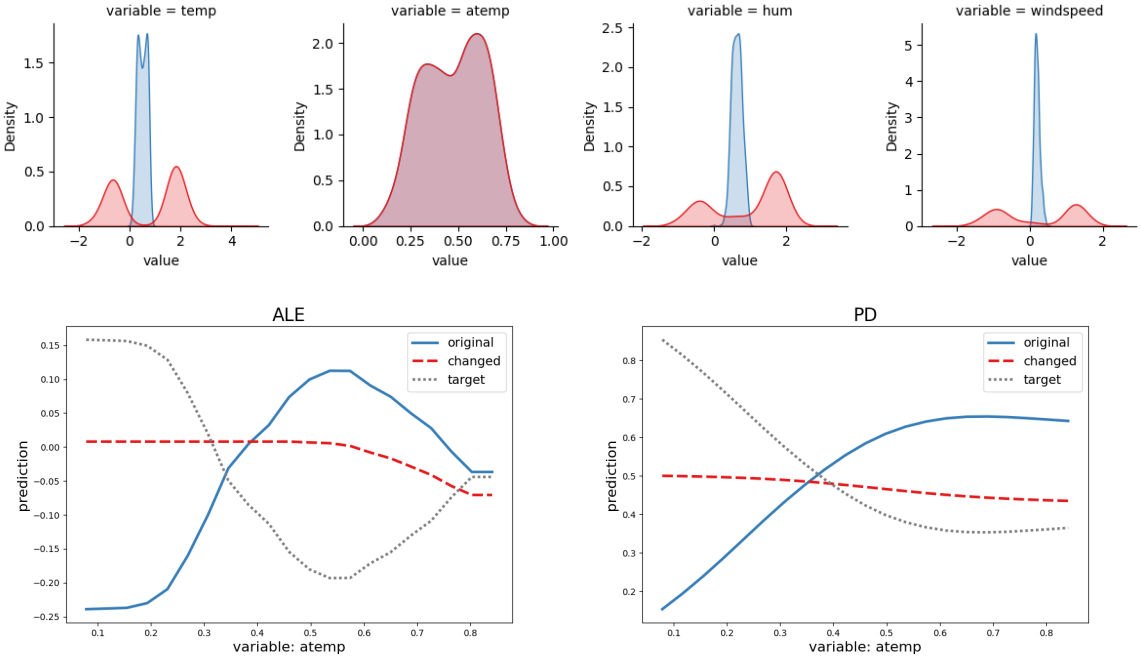


Figure 3: Plot showing the effect of `dist_weight` on poisoning results for different datasets.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.1

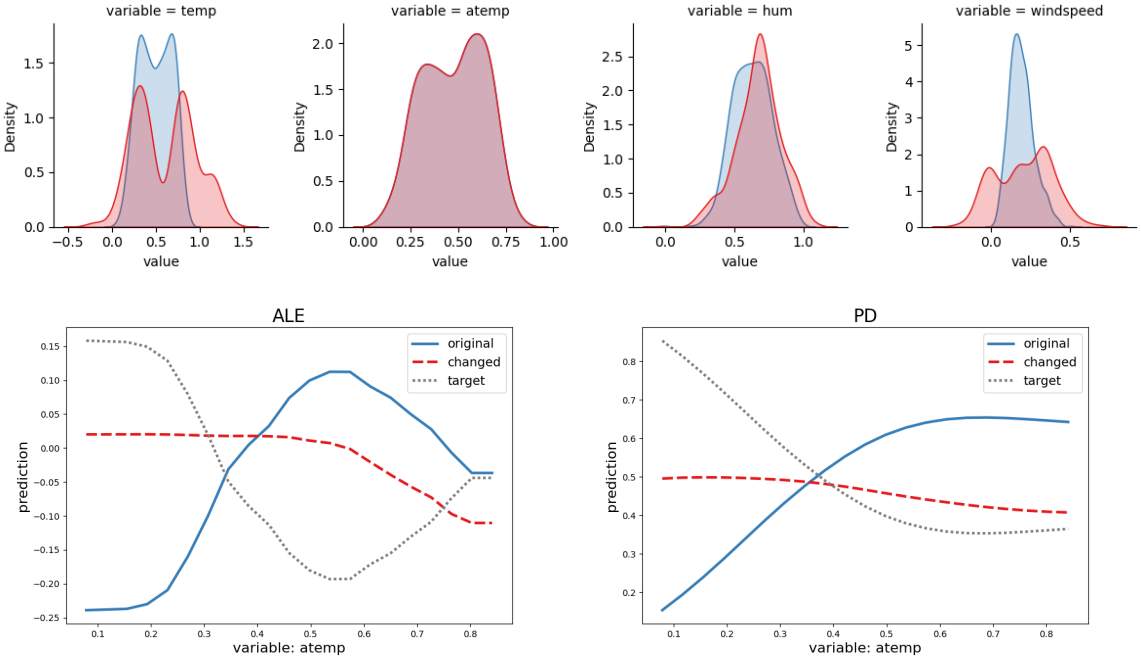
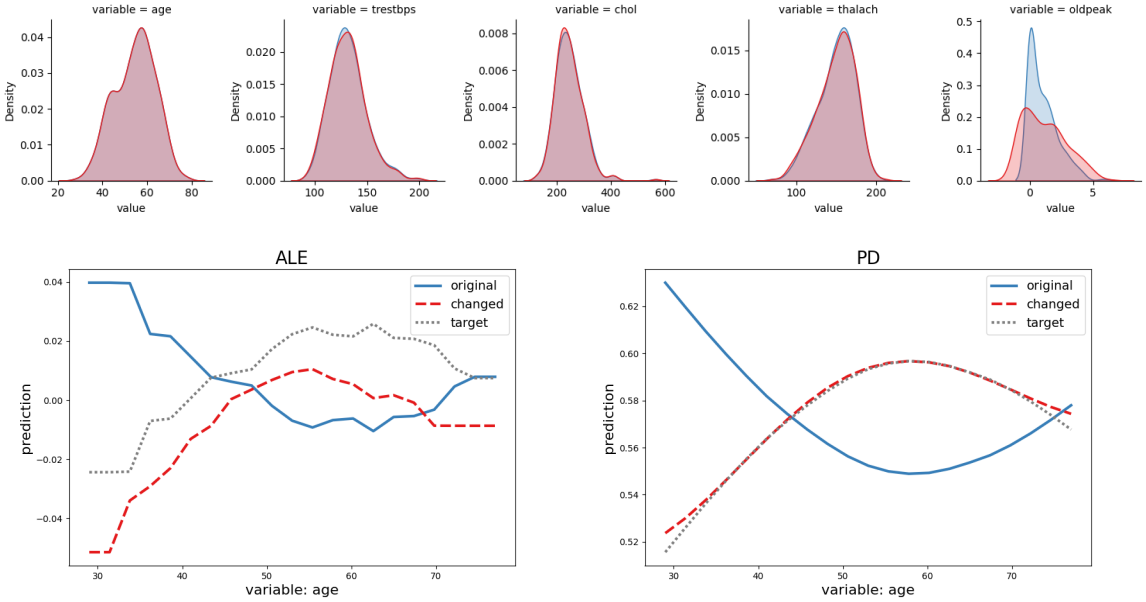


Figure 4: Comparison of poisoning results for the **bike** dataset, variable atemp.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.001

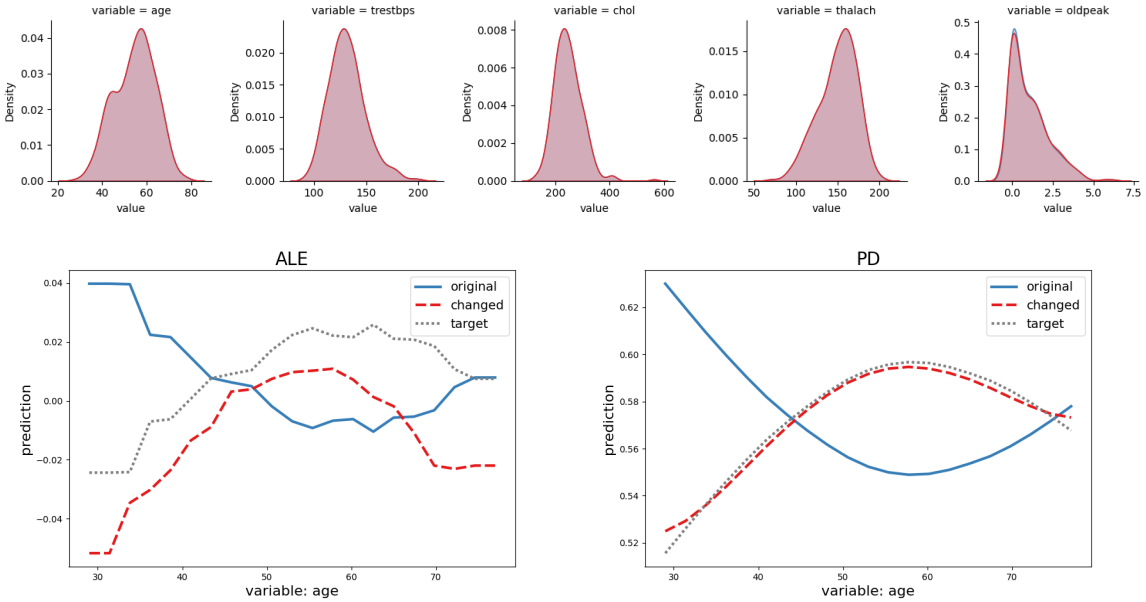
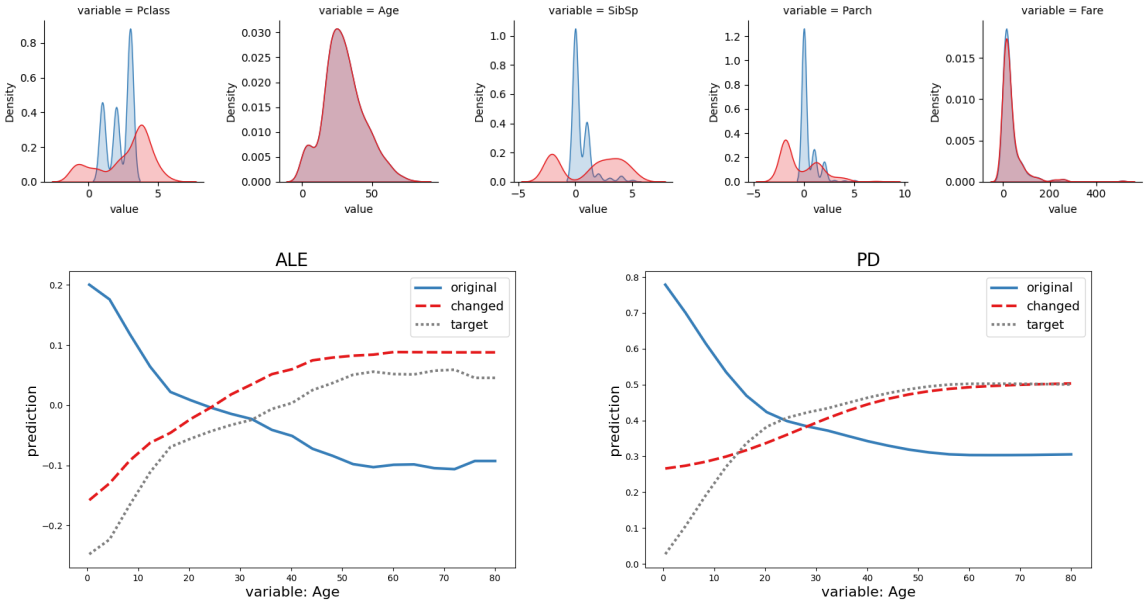


Figure 5: Comparison of poisoning results for the **heart** dataset, variable age.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.01

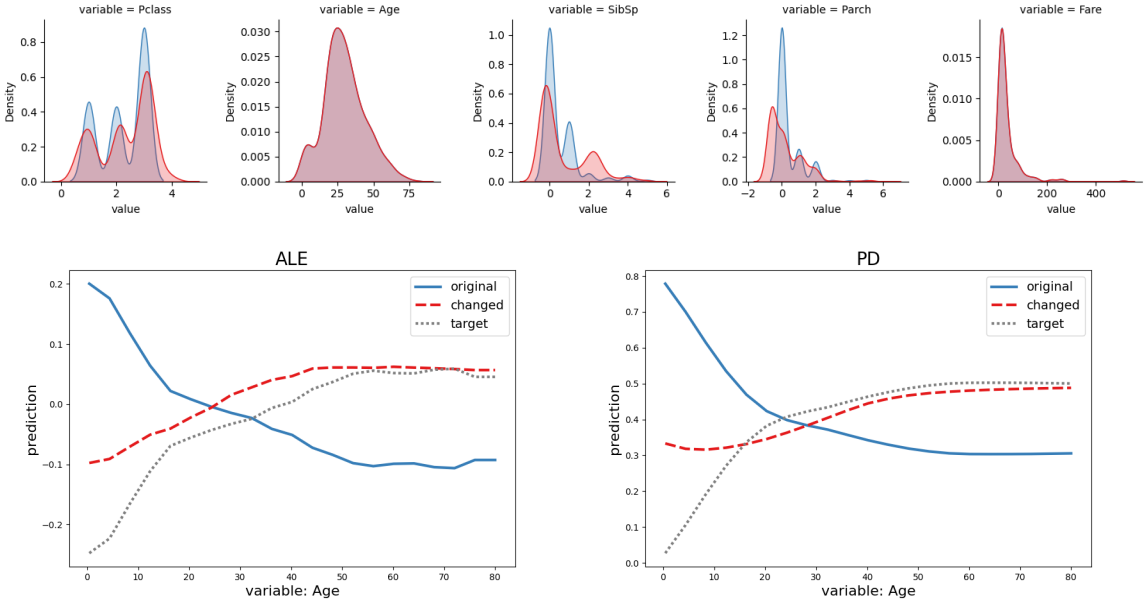
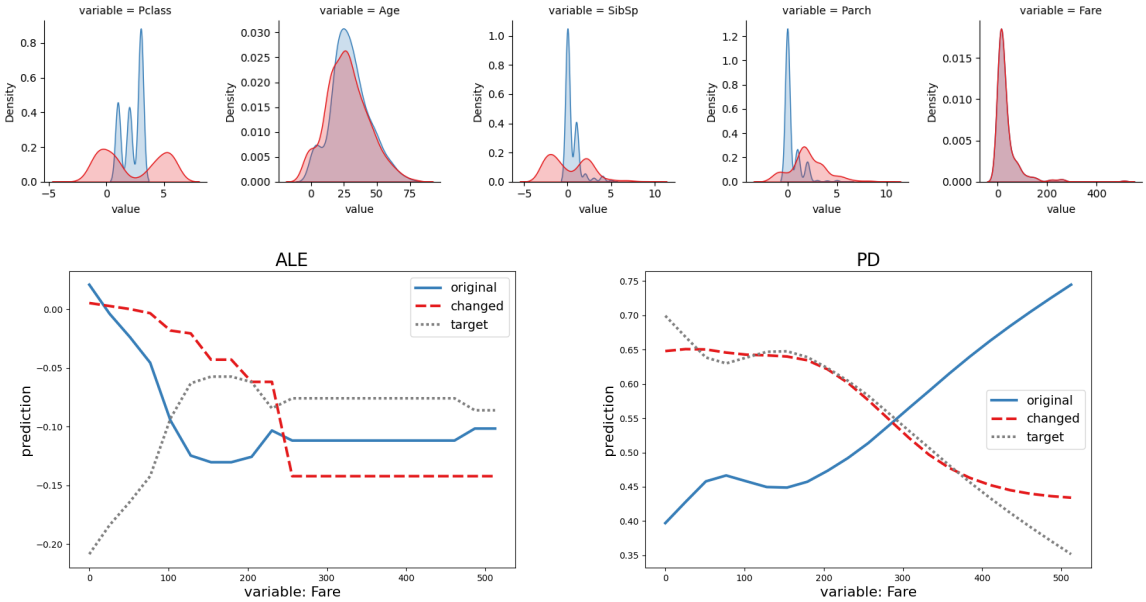


Figure 6: Comparison of poisoning results for the **titanic** dataset, variable Age.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.01

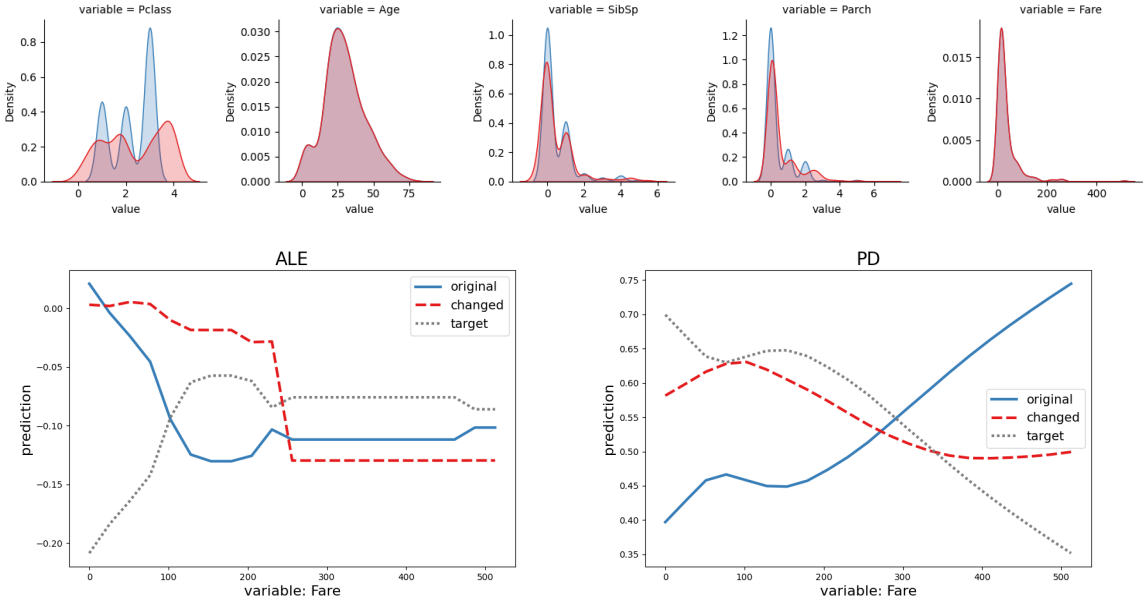
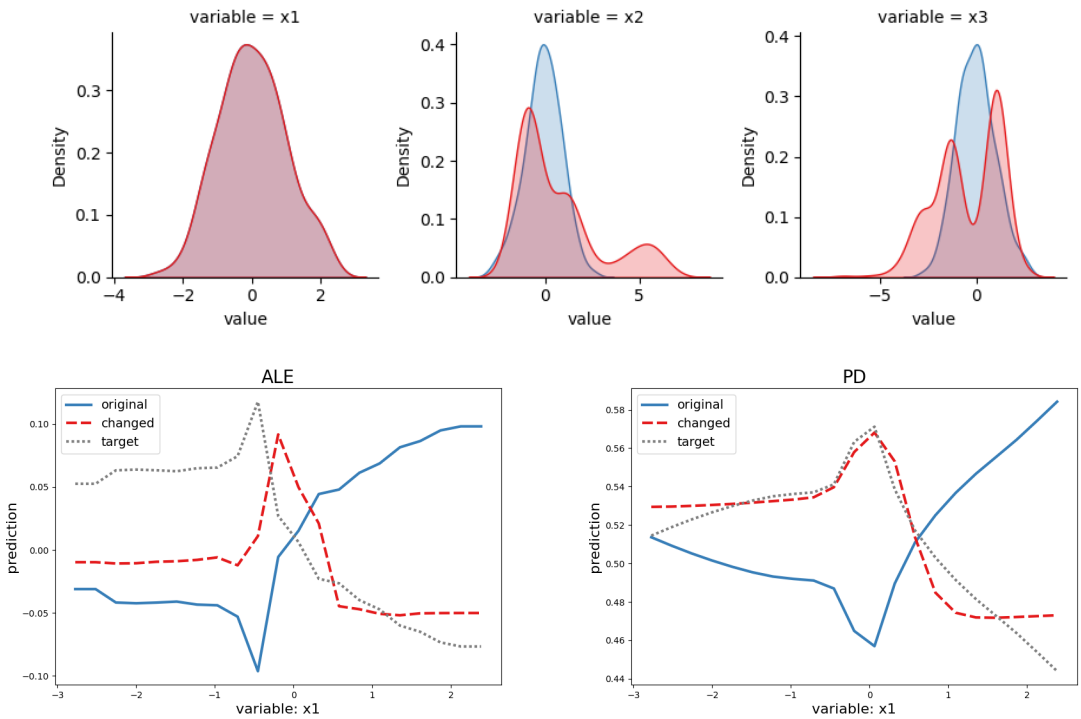


Figure 7: Comparison of poisoning results for the **titanic** dataset, variable Fare.

(a) Explanations and data distributions for `dist_weight` = 0



(b) Explanations and data distributions for `dist_weight` = 0.01

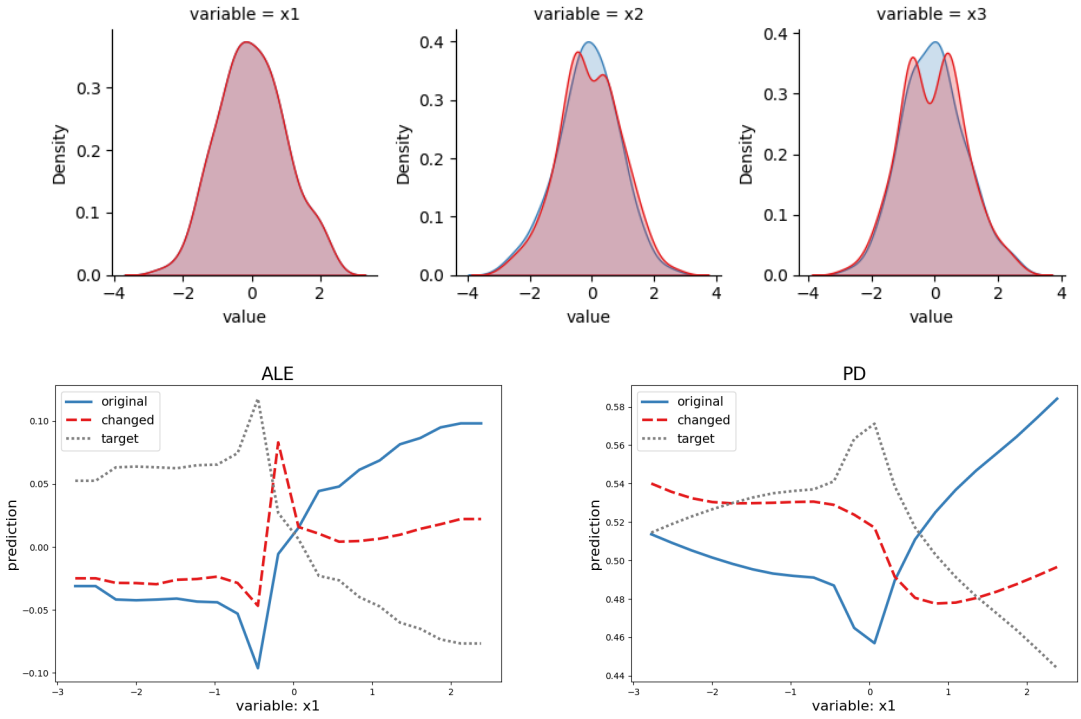


Figure 8: Comparison of poisoning results for the **xor** dataset, variable x1.