Robust and Stable Black Box Explanations

Himabindu Lakkaraju [1] **Nino Arsov** [2] **Osbert Bastani** [3]

A. Additional Results

A.1. Robustness to Real Distribution Shifts

We assess the robustness of explanations constructed using our approaches and the baselines on various real world datasets. The analysis that we present here is the same as that in Section 4.2, except for the underlying black boxes. In particular, we consider gradient boosted trees, random forests, and SVMs as black boxes. Corresponding results are presented in Tables 1, 2, and 3 respectively.

We observe similar results as that of Section 4.2 with other black boxes. All the explanations constructed using our framework ROPE have a much smaller drop in fidelity (0% to 5%) compared to those generated using the baselines. These results demonstrate that our approach significantly improves robustness. MUSE explanations have the largest percentage drop (13% to 26%). In contrast, both LIME and SHAP employ input perturbations when constructing explanations (**??**), resulting in somewhat increased robustness compared to MUSE. Nevertheless, LIME and SHAP still demonstrate a considerable drop, so they are still not very robust. Thus, these results validate our approach.

Tables 1, 2, and 3 also show the fidelities on both training data and shifted data. The fidelities of ROPE logistic and ROPE dset are lower than the other approaches, which is expected since ROPE logistic and ROPE dset only use a single logistic regression and a single decision set, respectively, to approximate the entire black box. On the other hand, ROPE logistic multi and ROPE dset multi achieve fidelities that are equal or better than the other baselines. These results demonstrate that ROPE achieves robustness without sacrificing fidelity on the original training distribution. Thus, our approach *strictly outperforms* the baseline approaches.

A.2. Impact of Degree of Distribution Shift on Fidelity

We replicate the analysis in Section 4.3, but with different black boxes. In particular, we consider gradient boosted trees, random forests, and SVMs as black boxes. Results are shown in Figures 1, 2, and 3, respectively. We observe similar patterns and trends as in Section 4.3.

[1]Harvard University [2]Macedonian Academy of Arts & Sciences [3]University of Pennsylvania. Correspondence to: Himabindu Lakkaraju <hlakkaraju@hbs.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Algorithms	**Bail**			**Academic**			**Health**		
	Train	Shift	% Drop	Train	Shift	% Drop	Train	Shift	% Drop
LIME	0.73	0.61	16.31%	0.71	0.59	17.38%	0.78	0.67	14.31%
SHAP	0.72	0.61	15.72%	0.69	0.58	16.37%	0.79	0.68	13.92%
MUSE	0.69	0.57	18.02%	0.67	0.53	20.32%	0.75	0.62	17.01%
ROPE logistic	0.59	0.57	3.02%	0.57	0.55	3.57%	0.68	0.66	2.32%
ROPE dset	0.63	0.61	2.98%	0.61	0.59	3.52%	0.74	0.73	1.92%
ROPE logistic multi	0.74	0.72	2.28%	0.71	0.69	2.45%	0.82	0.80	1.90%
ROPE dset multi	**0.76**	**0.74**	**2.13%**	**0.72**	**0.71**	**1.98%**	**0.83**	**0.81**	**1.89%**

Table 1. Gradient Boosted Trees (100 trees) as the black box. Fidelity values of all the explanations are reported on both training data and shifted data, along with percentage drop in fidelity from training data to shifted data. Smaller values of percentage drop correspond to more robust explanations.

Algorithms	**Bail**			**Academic**			**Health**		
	Train	Shift	% Drop	Train	Shift	% Drop	Train	Shift	% Drop
LIME	0.77	0.66	14.38%	0.69	0.61	11.83%	0.79	0.70	10.83%
SHAP	0.74	0.61	16.98%	0.67	0.58	12.82%	0.77	0.69	11.02%
MUSE	0.72	0.58	19.02%	0.65	0.55	15.01%	0.74	0.64	13.93%
ROPE logistic	0.63	0.62	2.32%	0.61	0.60	1.64%	0.69	0.68	**1.59%**
ROPE dset	0.65	0.64	1.97%	0.63	0.62	**1.02%**	0.70	0.69	1.61%
ROPE logistic multi	0.78	0.76	2.38%	0.73	0.71	3.12%	0.83	0.81	2.83%
ROPE dset multi	**0.79**	**0.77**	**1.92%**	**0.77**	**0.75**	2.03%	**0.86**	**0.84**	1.77%

Table 2. Random Forests (100 trees) as the black box. Fidelity values of all the explanations are reported on both training data and shifted data, along with percentage drop in fidelity from training data to shifted data. Smaller values of percentage drop correspond to more robust explanations.

Algorithms	**Bail**			**Academic**			**Health**		
	Train	Shift	% Drop	Train	Shift	% Drop	Train	Shift	% Drop
LIME	0.87	0.71	18.32%	0.89	0.74	17.27%	0.93	0.75	19.28%
SHAP	0.87	0.73	16.32%	0.91	0.76	15.98%	0.93	0.79	15.56%
MUSE	0.86	0.64	25.32%	0.87	0.67	23.41%	0.88	0.69	21.08%
ROPE logistic	0.81	0.79	2.39%	0.84	0.83	**1.08%**	0.87	0.86	**0.98%**
ROPE dset	0.84	0.82	2.50%	0.86	0.84	2.32%	0.89	0.86	2.98%
ROPE logistic multi	0.89	0.87	**1.98%**	0.92	0.89	3.32%	0.95	0.91	3.92%
ROPE dset multi	**0.93**	**0.91**	2.08%	**0.93**	**0.90**	3.32%	**0.96**	**0.92**	4.31%

Table 3. SVM as the black box. Fidelity values of all the explanations are reported on both training data and shifted data, along with percentage drop in fidelity from training data to shifted data. Smaller values of percentage drop correspond to more robust explanations.

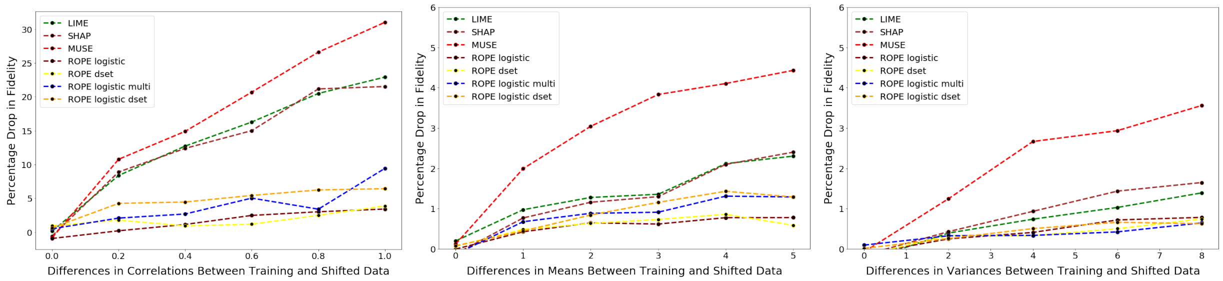

Figure 1. Gradient Boosted Trees (100 trees) as the black box. Impact of changes in covariate correlations (left), means (middle), and variances (right) on percentage drop in fidelities. Lower values of percentage drop indicate higher robustness. Standard errors are too small to be included.

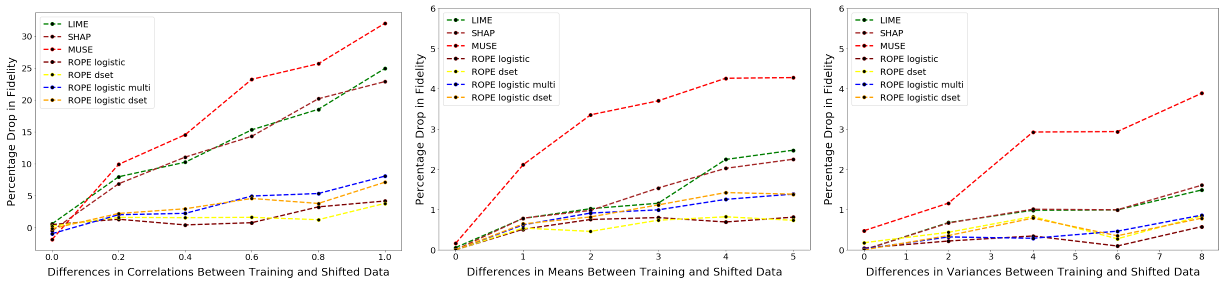

Figure 2. Random Forests (100 trees) as the black box. Impact of changes in covariate correlations (left), means (middle), and variances (right) on percentage drop in fidelities. Lower values of percentage drop indicate higher robustness. Standard errors are too small to be included.

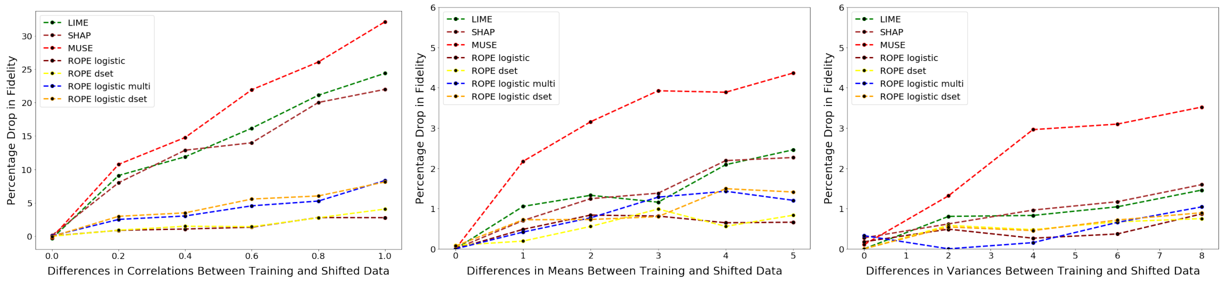

Figure 3. SVM as the black box. Impact of changes in covariate correlations (left), means (middle), and variances (right) on percentage drop in fidelities. Lower values of percentage drop indicate higher robustness. Standard errors are too small to be included.