Reproducibility Report for ACM SIGMOD 2024 Paper: "FACET: Robust Counterfactual Explanation Analytics"

Peter M. VanNostrand
pvannostrand@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

Konstantinos Kanellis
kkanellis@cs.wisc.edu
University of Wisconsin-Madison
Madison, USA

Wan Shen Lim
wanshenl@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Donghyun Sohn
donghyun.sohn@u.northwestern.edu
Northwestern University
Evanston, USA

Abstract

We are able to faithfully reproduce the original paper's findings as well as the key results reported in its experimental section (i.e., explanation quality, speed, robustness). The authors provided example data, comprehensive scripts, and plotting functions that allowed near-identical reconstruction of all of the paper's figures.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Explainable AI, Counterfactual Explanation, Random Forest, Gradient Boosting Ensembles, Interpretable Machine Learning.

ACM Reference Format:
Peter M. VanNostrand, Konstantinos Kanellis, Wan Shen Lim, and Donghyun Sohn. 2024. Reproducibility Report for ACM SIGMOD 2024 Paper: "FACET: Robust Counterfactual Explanation Analytics". In *Reproducibility Reports of the 2024 International Conference on Management of Data (SIGMOD ARI Reports '24), June 9–15, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3687998.3717053

1 Introduction

FACET [1] provides an explanation analytics system that supports the interactive refining of counterfactual explanations for decisions made by tree ensembles. The authors introduce the idea of a *counterfactual region* to concisely describe areas of the feature space in which the desired outcome is guaranteed, regardless of feature value variations.

The authors provide a script that allow for near-identical reconstruction of the paper's figures; each figure can be reproduced with an associated flag.

SIGMOD ARI Reports '24, Santiago, AA, Chile

ACM ISBN 979-8-4007-1179-4/2024/06
https://doi.org/10.1145/3687998.3717053

2 Submission

The paper's artifact includes a reproducibility script to run all experiments and generate all figures with a single command. A separate reproducibility README contains explicit and comprehensive instructions for using the script:

- Code Repository: https://github.com/PeterVanNostrand/FACET
- README: https://github.com/PeterVanNostrand/FACET/blob/main/instructions/REPRODUCABILITY.md
- The README provides instructions for setting up an Anaconda based Python environment to run the experiments. It also includes optional setup steps for Gurobi to enable comparison against an existing state-of-the-art method. After the environment is ready, a single command runs all experiments and generates all figures: `python replicate_paper.py --all_results --all_iterations`

3 Hardware and Software Environment

We provide a comparison of the machine setups for the authors and reproducers in Table 1. The README estimates a run time of approximately 12.5 hours for scaled-down experiments and one week for the full experiments, which matches what we observed.

4 Reproducibility Evaluation

4.1 Process

As described in the README, we set up an Anaconda environment and supplied an academic license key for Gurobi. We used a multiplexing terminal emulator (`tmux`) to run the reproducibility Python script to completion over the course of approximately one week – no attention or intervention was required during the entire run. If readers wish for fine-grained control, the reproducibility script supports flags for each experiment and figure (e.g., `--fig11 --fig12` to target figure 11 and figure 12).

On completion, the reproducibility script stores the raw result data in `results/` and the generated figures in `figures/`. As part of the reproducibility process, we discovered minor differences in the ensemble depth hyperparameter between the original paper submission and final code repository. Subsequently, the authors have updated the reproducibility script to run with both hyperparameter settings and store the resulting plots in `figures/final-github` and `figures/final-paper` respectively. Future work should use the code repository's default hyperparameters.

Table 1: Hardware & Software environment

	Paper	Reviewer A	Reviewer B	Reviewer C
CPU	AMD EPYC-7543	Intel(R) Xeon Silver 4114	Intel(R) Xeon(R) Gold 5218R	Intel Core i7-9750H
Cores	4 cores (× 1 threads)	10 (× 2 threads)	2 × 20 (× 2 threads)	6 (× 2 threads)
GHz	2.8 GHz	2.2 GHz	2.10 GHz	2.6 GHz
RAM	16 GB	192 GB	180 GB	16 GB
Storage	VAST Scale-Out Storage	Intel(R) DC S3500 480 GB	SSD	500 GB SSD
Operating System	Ubuntu 20.04 LTS	Ubuntu 22.04 LTS	Ubuntu 22.04.2 LTS	macOS 15.0.1

4.2 Results

The reproducibility script successfully ran all experiments and generated all the figures in the paper. In this section, we highlight Table 3 and Figure 9 of the original paper [1]. Table 3 shows that FACET beats existing state-of-the-art methods in explanation quality and speed, whereas Figure 9 demonstrates FACET's robustness. We show an excerpt of our reproduced results in Figures 1 and 2.

5 Summary

All experimental results are reproducible on our machines. The reviewers commend the authors for the ease of reproduction, with excerpts from our private discussions below:

- The authors provide very clear instructions and automated scripts for all steps.
- The author was highly responsive to email inquiries. It is clear that they have put significant effort into making it easy for others to reproduce their figures.
- The current implementation, with its automated experiment suite and clear separation of setup and execution phases, serves as an excellent example of how artifacts should be organized.

References

[1] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. 2023. FACET: Robust counterfactual explanation analytics. *Proceedings of the ACM on Management of Data* 1, 4 (2023), 1–27.

Dataset	ADULT					CANCER*					CREDIT					MAGIC					SPAMBASE*				
	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑
FACET	0.108	1.68	1.00	1.00	100	0.094	4.33	1.00	1.00	100	0.128	4.54	1.00	1.00	100	0.003	2.99	1.00	1.00	100	0.201	4.47	1.00	1.00	100
MACE	107.4	5.69	70.5	81.4	68.5	2.222	30.00	21.2	8.88	100	190.2	10.9	9.07	5.77	38.0	2.472	10.0	12.4	7.61	100	19.91	57.0	1261	371	100
OCEAN	0.586	2.44	3.26	2.97	100	0.681	11.41	1.43	1.02	100	1.721	6.15	2.13	1.97	100	1.790	6.30	1.45	1.05	100	0.464	7.50	0.69	0.64	100
RFOCSE	56.59	1.40	0.54	0.58	90.5	1936	4.65	1.29	1.07	100	202.5	2.44	0.16	0.21	35.0	147.0	2.42	0.87	1.04	95.0	2208	3.30	4.25	4.51	100
AFT	0.003	1.46	0.76	0.80	96.0	0.002	1.86	0.92	1.26	37.0	0.002	2.42	1.12	1.22	85.5	0.004	1.65	0.99	1.24	99.0	0.003	1.98	2.00	3.03	76.0
FCT-RF	0.233	3.92	1.00	1.00	100	0.110	11.86	1.00	1.00	100	0.129	6.31	1.00	1.00	100	0.007	4.24	1.00	1.00	100	0.035	11.04	1.00	1.00	100
FCT-GB1	0.338	2.60	0.11	0.15	100	0.054	8.08	0.50	0.60	100	0.119	4.15	0.13	0.14	100	0.005	4.86	0.78	0.68	100	0.052	7.75	0.51	0.59	100
FCT-GB2	0.299	2.48	0.08	0.12	100	0.047	7.52	0.47	0.58	100	0.118	4.09	0.12	0.13	100	0.005	4.83	0.77	0.68	100	0.050	7.61	0.50	0.55	100

(a) Paper

Dataset	ADULT					CANCER*					CREDIT					MAGIC					SPAMBASE*				
	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑
FACET	0.108	1.68	1.00	1.00	100	0.094	4.33	1.00	1.00	100	0.128	4.54	1.00	1.00	100	0.003	2.99	1.00	1.00	100	0.201	4.47	1.00	1.00	100
MACE	107.4	5.69	70.5	81.4	68.5	2.222	30.00	21.2	8.88	100	190.2	10.9	9.07	5.77	38.0	2.472	10.0	12.4	7.61	100	19.91	57.0	1261	371	100
OCEAN	0.586	2.44	3.26	2.97	100	0.681	11.41	1.43	1.02	100	1.721	6.15	2.13	1.97	100	1.790	6.30	1.45	1.05	100	0.464	7.50	0.69	0.64	100
RFOCSE	56.59	1.40	0.54	0.58	90.5	1936	4.65	1.29	1.07	100	202.5	2.44	0.16	0.21	35.0	147.0	2.42	0.87	1.04	95.0	2208	3.30	4.25	4.51	100
AFT	0.003	1.46	0.76	0.80	96.0	0.002	1.86	0.92	1.26	37.0	0.002	2.42	1.12	1.22	85.5	0.004	1.65	0.99	1.24	99.0	0.003	1.98	2.00	3.03	76.0
FCT-RF	0.233	3.92	1.00	1.00	100	0.110	11.86	1.00	1.00	100	0.129	6.31	1.00	1.00	100	0.007	4.24	1.00	1.00	100	0.035	11.04	1.00	1.00	100
FCT-GB1	0.338	2.60	0.11	0.15	100	0.054	8.08	0.50	0.60	100	0.119	4.15	0.13	0.14	100	0.005	4.86	0.78	0.68	100	0.052	7.75	0.51	0.59	100
FCT-GB2	0.299	2.48	0.08	0.12	100	0.047	7.52	0.47	0.58	100	0.118	4.09	0.12	0.13	100	0.005	4.83	0.77	0.68	100	0.050	7.61	0.50	0.55	100

(b) Reviewer A

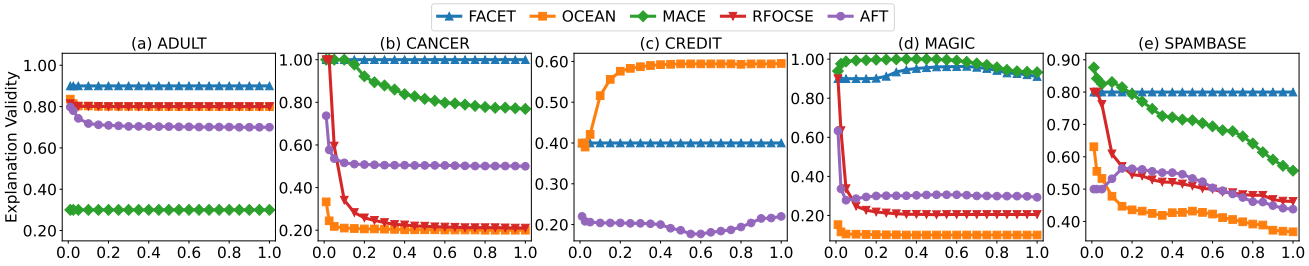
Dataset	ADULT					CANCER*					CREDIT					MAGIC					SPAMBASE*				
	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑
FACET	0.108	1.68	1.00	1.00	100	0.094	4.33	1.00	1.00	100	0.128	4.54	1.00	1.00	100	0.003	2.99	1.00	1.00	100	0.201	4.47	1.00	1.00	100
MACE	107.4	5.69	70.5	81.4	68.5	2.222	30.00	21.2	8.88	100	190.2	10.9	9.07	5.77	38.0	2.472	10.0	12.4	7.61	100	19.91	57.0	1261	371	100
OCEAN	0.586	2.44	3.26	2.97	100	0.681	11.41	1.43	1.02	100	1.721	6.15	2.13	1.97	100	1.790	6.30	1.45	1.05	100	0.464	7.50	0.69	0.64	100
RFOCSE	56.59	1.40	0.54	0.58	90.5	1936	4.65	1.29	1.07	100	202.5	2.44	0.16	0.21	35.0	147.0	2.42	0.87	1.04	95.0	2208	3.30	4.25	4.51	100
AFT	0.003	1.46	0.76	0.80	96.0	0.002	1.86	0.92	1.26	37.0	0.002	2.42	1.12	1.22	85.5	0.004	1.65	0.99	1.24	99.0	0.003	1.98	2.00	3.03	76.0
FCT-RF	0.233	3.92	1.00	1.00	100	0.110	11.86	1.00	1.00	100	0.129	6.31	1.00	1.00	100	0.007	4.24	1.00	1.00	100	0.035	11.04	1.00	1.00	100
FCT-GB1	0.338	2.60	0.11	0.15	100	0.054	8.08	0.50	0.60	100	0.119	4.15	0.13	0.14	100	0.005	4.86	0.78	0.68	100	0.052	7.75	0.51	0.59	100
FCT-GB2	0.299	2.48	0.08	0.12	100	0.047	7.52	0.47	0.58	100	0.118	4.09	0.12	0.13	100	0.005	4.83	0.77	0.68	100	0.050	7.61	0.50	0.55	100

(c) Reviewer B

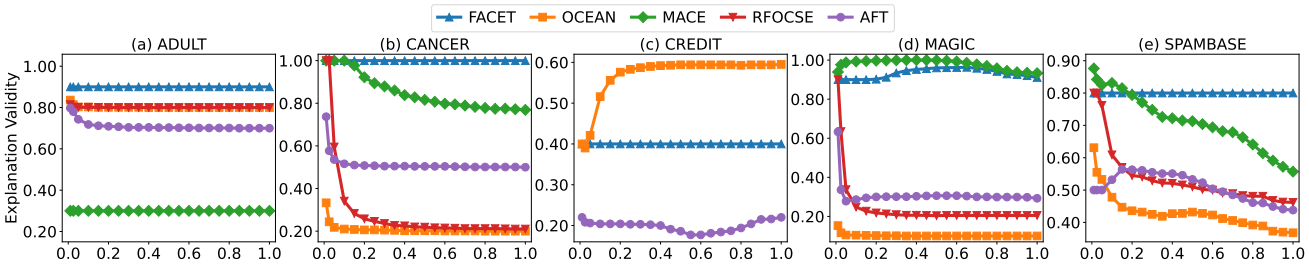
Dataset	ADULT					CANCER*					CREDIT					MAGIC					SPAMBASE*				
	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑	$t\downarrow$	$\delta_0\downarrow$	$\delta_1\downarrow$	$\delta_2\downarrow$	%↑
FACET	0.108	1.68	1.00	1.00	100	0.094	4.33	1.00	1.00	100	0.128	4.54	1.00	1.00	100	0.003	2.99	1.00	1.00	100	0.201	4.47	1.00	1.00	100
MACE	107.4	5.69	70.5	81.4	68.5	2.222	30.00	21.2	8.88	100	190.2	10.9	9.07	5.77	38.0	2.472	10.0	12.4	7.61	100	19.91	57.0	1261	371	100
OCEAN	0.586	2.44	3.26	2.97	100	0.681	11.41	1.43	1.02	100	1.721	6.15	2.13	1.97	100	1.790	6.30	1.45	1.05	100	0.464	7.50	0.69	0.64	100
RFOCSE	56.59	1.40	0.54	0.58	90.5	1936	4.65	1.29	1.07	100	202.5	2.44	0.16	0.21	35.0	147.0	2.42	0.87	1.04	95.0	2208	3.30	4.25	4.51	100
AFT	0.003	1.46	0.76	0.80	96.0	0.002	1.86	0.92	1.26	37.0	0.002	2.42	1.12	1.22	85.5	0.004	1.65	0.99	1.24	99.0	0.003	1.98	2.00	3.03	76.0
FCT-RF	0.233	3.92	1.00	1.00	100	0.110	11.86	1.00	1.00	100	0.129	6.31	1.00	1.00	100	0.007	4.24	1.00	1.00	100	0.035	11.04	1.00	1.00	100
FCT-GB1	0.338	2.60	0.11	0.15	100	0.054	8.08	0.50	0.60	100	0.119	4.15	0.13	0.14	100	0.005	4.86	0.78	0.68	100	0.052	7.75	0.51	0.59	100
FCT-GB2	0.299	2.48	0.08	0.12	100	0.047	7.52	0.47	0.58	100	0.118	4.09	0.12	0.13	100	0.005	4.83	0.77	0.68	100	0.050	7.61	0.50	0.55	100

(d) Reviewer C

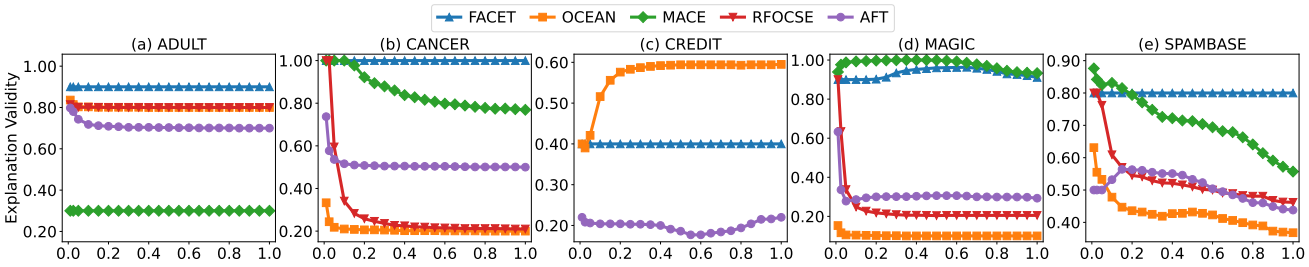
Figure 1: Reproduction (excerpt) of Table 3 in [1]: Comparison to state-of-the art counterfactual example generation techniques for random forest (T = 10, $D_{\max}$ = 5) in terms time t, sparsity s, L1-Norm δ_1, L2-Norm δ_2, and validity %.



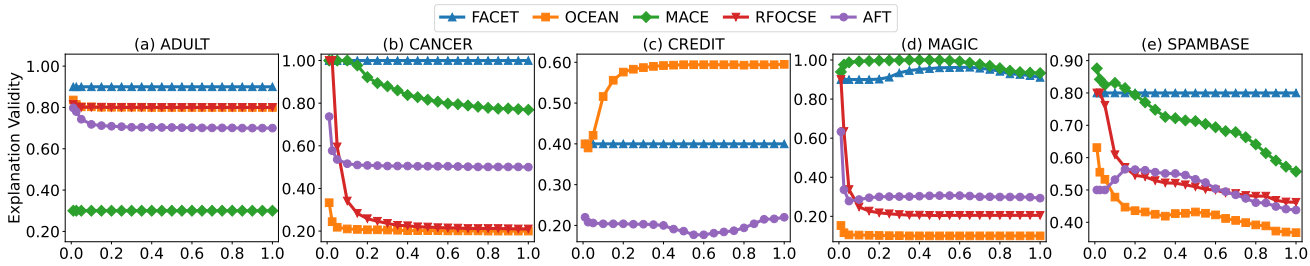
(a) Paper



(b) Reviewer A



(c) Reviewer B



(d) Reviewer C

Figure 2: Reproduction (excerpt) of Figure 9 in [1]: Evaluation of nearest explanation robustness to varying random perturbation size (percent of space).