Counterfactual Explanations

Srikar Katta
Ghazal Khalighinejad

Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.

Setup

- Predictive classifier f
- Instance $\boldsymbol{x}$ (observation), y (outcome)
- Example
  - $\boldsymbol{x}$: people
  - y: loan prediction

Counterfactual (CF) Explanations

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0

Counterfactual (CF) Explanations

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0
CF1	M	$100,000	Bachelor’s	…	1
CF2	M	$1,100,000	Bachelor’s	…	1
CF3	M	$100,000	Master’s	…	1

Counterfactual (CF) Explanations

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0
CF1	M	$100,000	Bachelor’s	…	1
CF2	M	$1,100,000	Bachelor’s	…	1
CF3	M	$100,000	Master’s	…	1

Question: what are the flaws of these explanations?

Counterfactual (CF) Explanations

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor's	…	0
CF1	M	$100,000	Bachelor's	…	1
CF2	M	$1,100,000	Bachelor's	…	1
CF3	M	$100,000	Master's	…	1
CF4	F	$110,000	Master's	…	1

What if we also saw CF4?

How would we solve this problem?

Setup

- Predictive classifier f
- Instance $\boldsymbol{x}$ (observation), y (outcome)

Setup

- Predictive classifier *f*
- Instance ***x*** (observation), *y* (outcome)
- Goal: create counterfactuals {c1, …, ck} that are
  - Diverse : different from one another

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0
Bad CF	M	$100,000	Bachelor’s	…	1
Good CF	F	$100,100	Bachelor’s	…	1

Setup

- Predictive classifier *f*
- Instance ***x*** (observation), *y* (outcome)
- Goal: create counterfactuals {c1, …, ck} that are
  - Proximal : close to the original instance

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0
Bad CF	M	$1,100,000	Bachelor’s	…	1
Good CF	F	$100,100	Bachelor’s	…	1

Setup

- Predictive classifier *f*
- Instance ***x*** (observation), *y* (outcome)
- Goal: create counterfactuals {c1, …, ck} that are
  - Sparse : do not involve too many features

	Gender	Income	Education	…	Loan prediction
Query unit	F	$100,000	Bachelor’s	…	0
Bad CF	M	$100,100	Master’s	…	1
Good CF	F	$100,100	Bachelor’s	…	1

Optimization

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \mathrm{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2\, \mathrm{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Optimization

$$C(\boldsymbol{x}) = \boxed{\underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min}} \frac{1}{k}\sum_{i=1}^{k} \mathrm{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2 \,\mathrm{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Find the *k* counterfactuals that minimize the following objective

Optimization

Check loss between the true outcome and the predicted outcome *given the counterfactual*

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k} \sum_{i=1}^{k} \boxed{\text{yloss}(f(\boldsymbol{c}_i), y)} + \frac{\lambda_1}{k} \sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2 \, \text{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Optimization

Check distance between the counterfactual and the given instance

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \mathrm{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} \boxed{dist(\boldsymbol{c}_i, \boldsymbol{x})} - \lambda_2\, \mathrm{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Optimization

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k} \sum_{i=1}^{k} \text{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x})$$
$$\boxed{- \lambda_2 \, \text{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)}$$

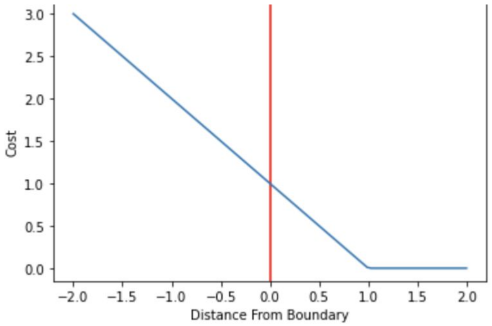
Increase how different counterfactuals are from one another

Practical considerations

- What should *yloss* be?
  - A valid counterfactual only needs to change the prediction to pass some threshold
  - Don't need to make prediction 0.49 → 0.99
  - Make a prediction of 0.49 → 0.51

Practical considerations

- What should *yloss* be?
  - A valid counterfactual only needs to change the prediction to pass some threshold
  - Don’t need to make prediction 0.49 → 0.99
  - Make a prediction of 0.49 → 0.51



Hinge loss

Practical considerations

- What should *yloss* be?
  - A valid counterfactual only needs to change the prediction to pass some threshold
  - Don’t need to make prediction 0.49 → 0.99
  - Make a prediction of 0.49 → 0.51
- What should distance be?

$$\text{dist\_cont}(\boldsymbol{c}, \boldsymbol{x}) = \frac{1}{d_{cont}} \sum_{p=1}^{d_{cont}} \frac{|\boldsymbol{c}^p - \boldsymbol{x}^p|}{MAD_p}$$

Practical considerations

- What should *yloss* be?
    - A valid counterfactual only needs to change the prediction to pass some threshold
    - Don’t need to make prediction 0.49 → 0.99
    - Make a prediction of 0.49 → 0.51
- What should distance be?

$$\text{dist\_cont}(\boldsymbol{c}, \boldsymbol{x}) = \frac{1}{d_{cont}} \sum_{p=1}^{d_{cont}} \frac{|\boldsymbol{c}^p - \boldsymbol{x}^p|}{MAD_p} \qquad \text{dist\_cat}(\boldsymbol{c}, \boldsymbol{x}) = \frac{1}{d_{cat}} \sum_{p=1}^{d_{cat}} I(\boldsymbol{c}^p \neq \boldsymbol{x}^p),$$

Practical considerations

- What should *yloss* be?
  - A valid counterfactual only needs to change the prediction to pass some threshold
  - Make a prediction of 0.49 --> 0.51, not 0.49 --> 0.99
- What should distance be?
- How do we induce sparsity?
  - Post-hoc, greedy approach
  - Keep adding values of cont. features back in until predicted class change

Sparsity Example

	Gender	Income	Education	…	Loan prediction
Query unit	F	\$100,000	Bachelor's	…	0
Original CF	M	\$1,100,000	Master's	…	1
Iteration 1	M	\$1,100,000	Bachelor's	…	1
Iteration 2	M	\$100,000	Bachelor's	…	1

How should we evaluate counterfactuals?

- *Validity*: the counterfactuals' predicted outcome is different than original outcome

How should we evaluate counterfactuals?

- *Validity*: the counterfactuals' predicted outcome is different than original outcome
- *Proximity*: the counterfactuals should be similar to the query instance

How should we evaluate counterfactuals?

- *Validity*: the counterfactuals' predicted outcome is different than original outcome
- *Proximity*: the counterfactuals should be similar to the query instance
- *Sparsity*: the counterfactuals should not require changing too many covariates

How should we evaluate counterfactuals?

- *Validity*: the counterfactuals' predicted outcome is different than original outcome
- *Proximity*: the counterfactuals should be similar to the query instance
- *Sparsity*: the counterfactuals should not require changing too many covariates
- *Diversity*: the counterfactuals should be different from one another

Experiments

- Baseline methods for explaining non-linear models
  - SingleCF

  Wachter’s algorithm – ours but without diversity term and only one counterfactual

Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF

Wachter's algorithm with *k* random starting points for optimizer

Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF
  - NoDiversityCF

Our algorithm but with *multiple* counterfactuals and no diversity term

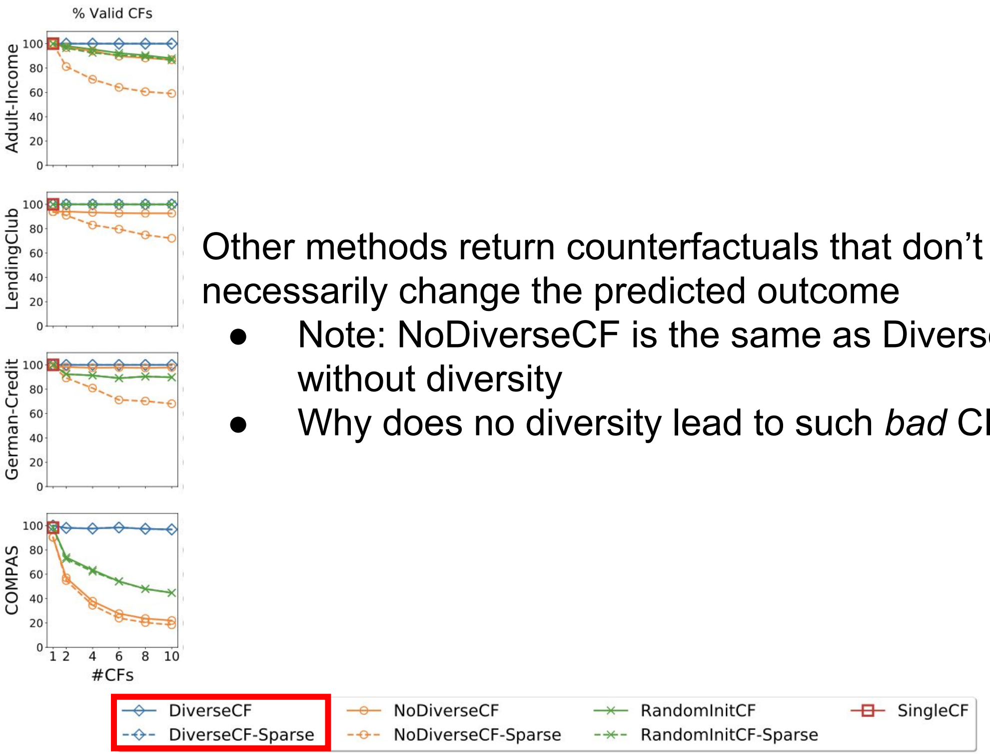
Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF
  - NoDiversityCF
- Baseline methods for explaining linear models
  - MixedIntegerCF

Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF
  - NoDiversityCF
- Baseline methods for explaining linear models
  - MixedIntegerCF
- Datasets
  - Adult income: Classify whether adult's income is over $50,000
  - COMPAS: Classify whether criminals will re-offend
  - German credit: Determine whether person has good/bad credit
  - LendingClub: Determine whether person will pay loan back or not

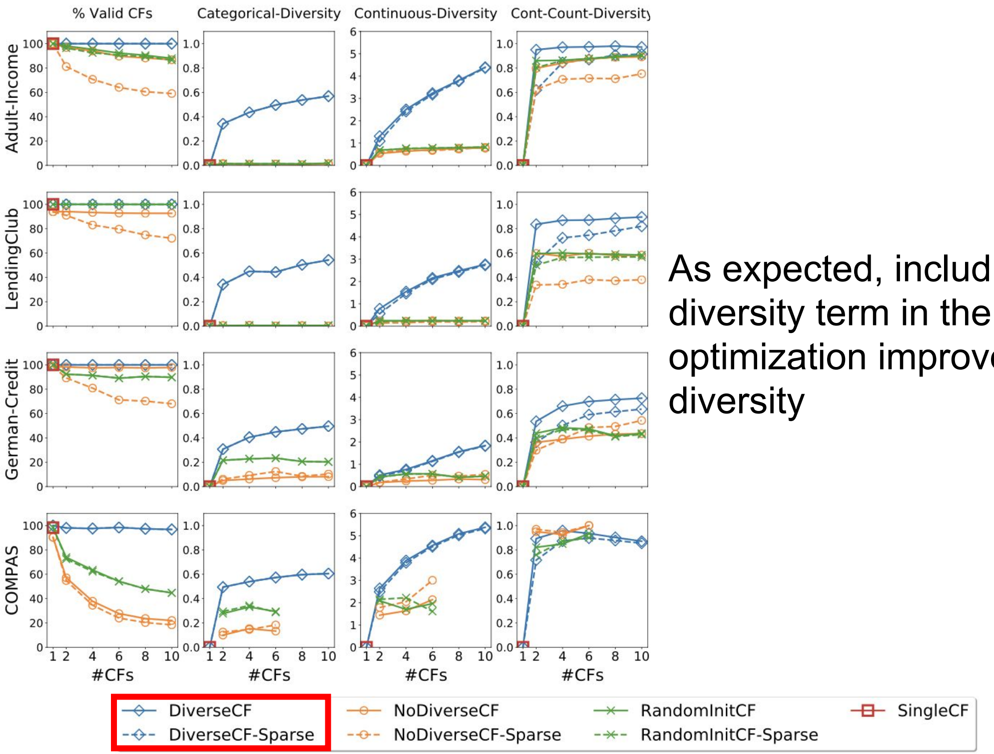
Explaining Non-linear Models



Other methods return counterfactuals that don't necessarily change the predicted outcome

- Note: NoDiverseCF is the same as DiverseCF but without diversity
- Why does no diversity lead to such *bad* CFs?

Explaining Non-linear Models

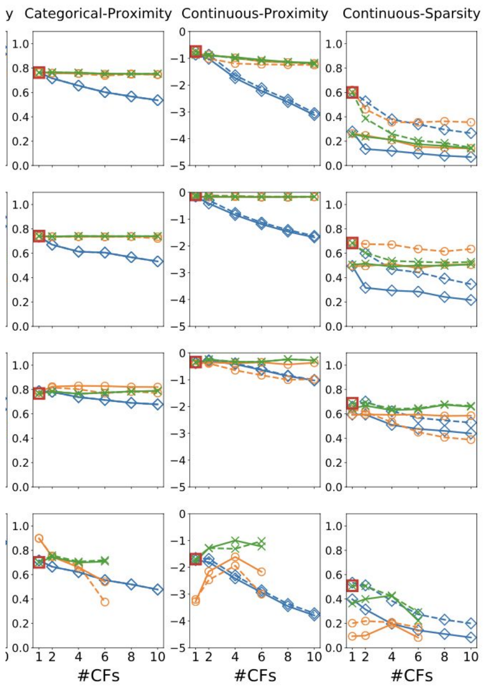


As expected, including a diversity term in the optimization improves diversity

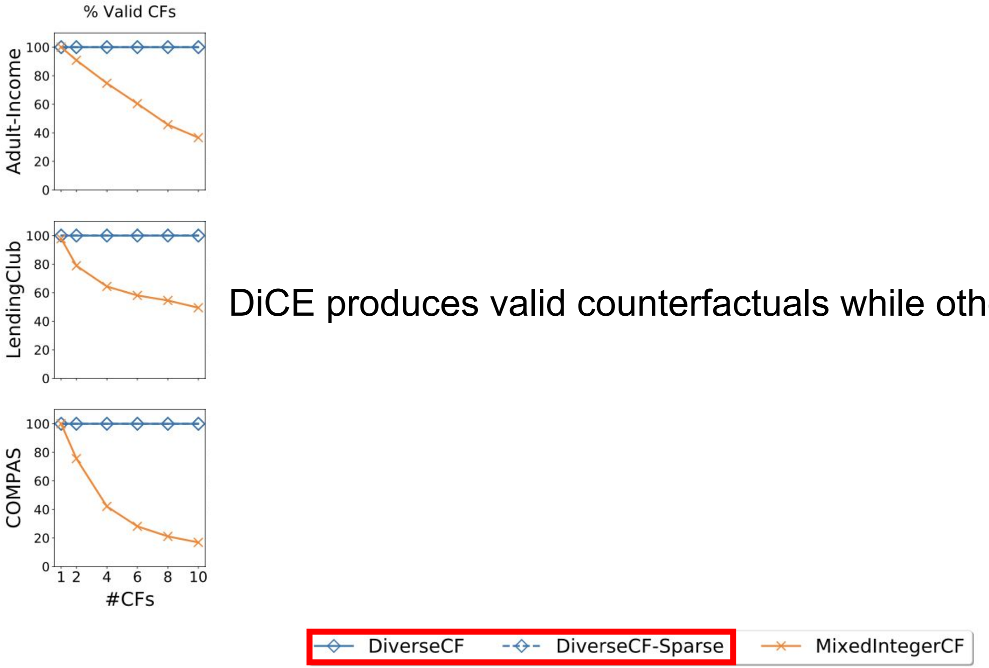
Explaining Non-linear Models

DiCE’s counterfactuals are more similar to the query unit than baseline methods

And the results are either as sparse (if not more) than other methods

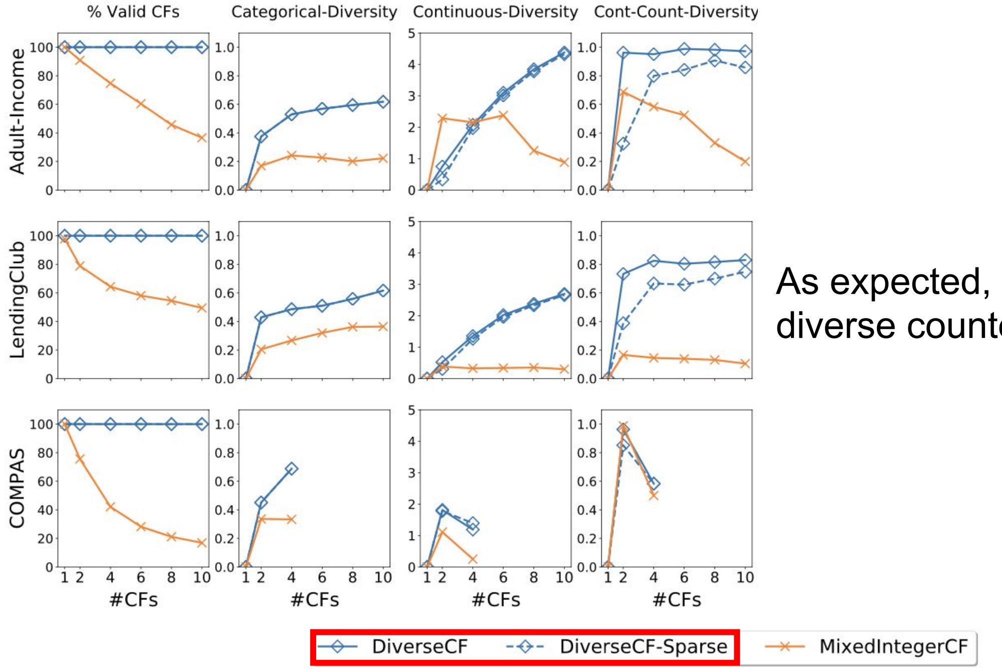


Explaining Linear Models



DiCE produces valid counterfactuals while others don't

Explaining Linear Models

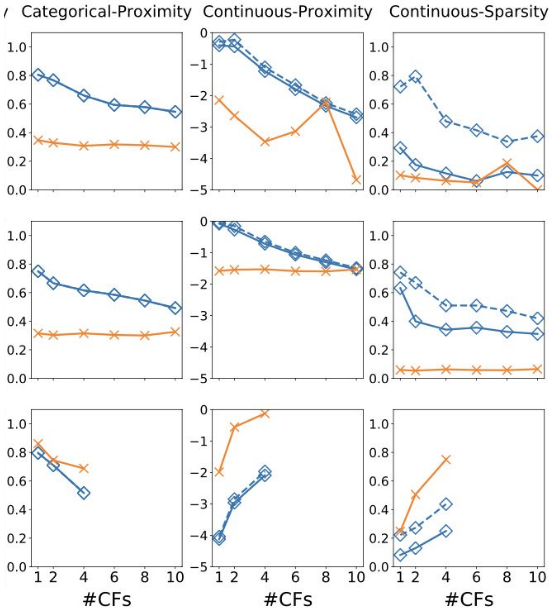


As expected, DiCE has more diverse counterfactuals

Explaining Linear Models

But DiCE's counterfactuals are not as close to the original instance

But DiCE is more sparse (for the most part



Qualitative Evaluation

Adult	HrsWk	Education	Occupation	WorkClass	Race	AgeYrs	MaritalStat	Sex
Original input (outcome: <=50K)	45.0	HS-grad	Service	Private	White	22.0	Single	Female
Counterfactuals (outcome: >50K)	—	Masters	—	—	—	65.0	Married	Male
	—	Doctorate	—	Self-Employed	—	34.0	—	—
	33.0	—	White-Collar	—	—	47.0	Married	—
	57.0	Prof-school	—	—	—	—	Married	—

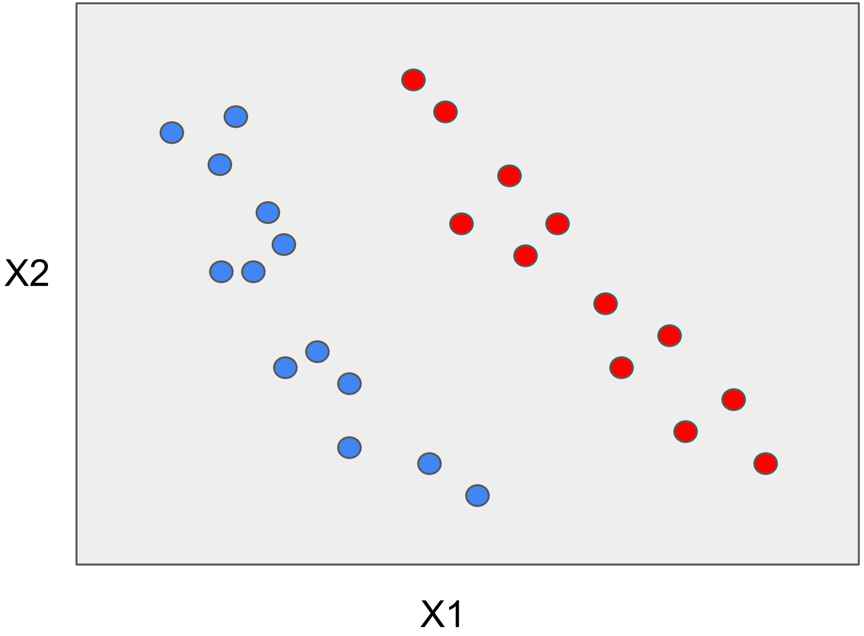
Qualitative Evaluation

LendingClub	EmpYrs	Inc$	#Ac	CrYrs	LoanGrade	HomeOwner	Purpose	State
Original input (outcome: Default)	7.0	69996.0	4.0	26.0	D	Mortgage	Debt	NY
Counterfactuals (outcome: Paid)	—	61477.0	—	—	B	—	Purchase	—
	10.0	83280.0	1.0	23.0	A	—	—	TX
	10.0	69798.0	—	40.0	A	—	—	—
	10.0	130572.0	—	—	A	Rent	—	—

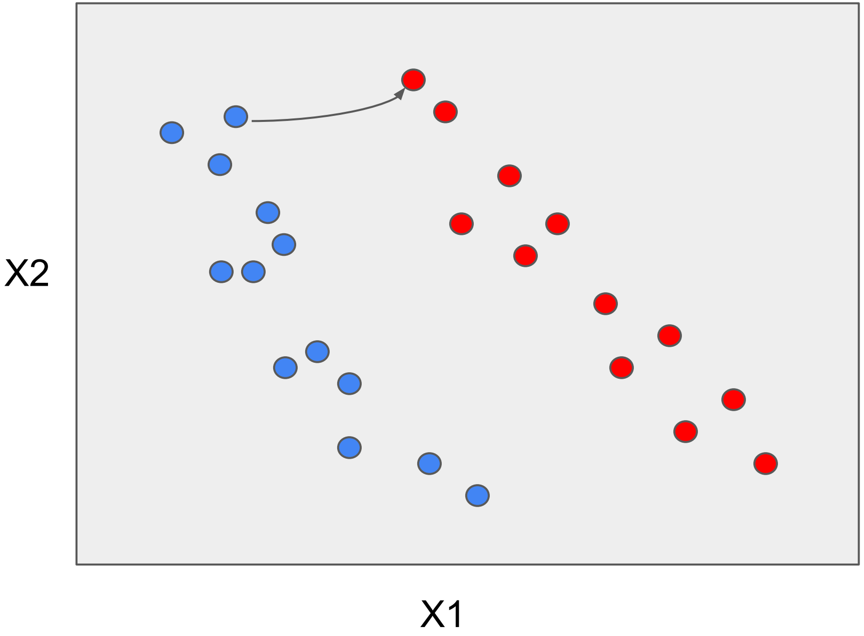
Qualitative Evaluation

COMPAS	PriorsCount	CrimeDegree	Race	Age	Sex
Original input (outcome: Will Recidivate)	10.0	Felony	African-American	>45	Female
Counterfactuals (outcome: Won't Recidivate)	—	—	Caucasian	—	—
	0.0	—	—	—	Male
	0.0	—	Hispanic	—	—
	9.0	Misdemeanor	—	—	—

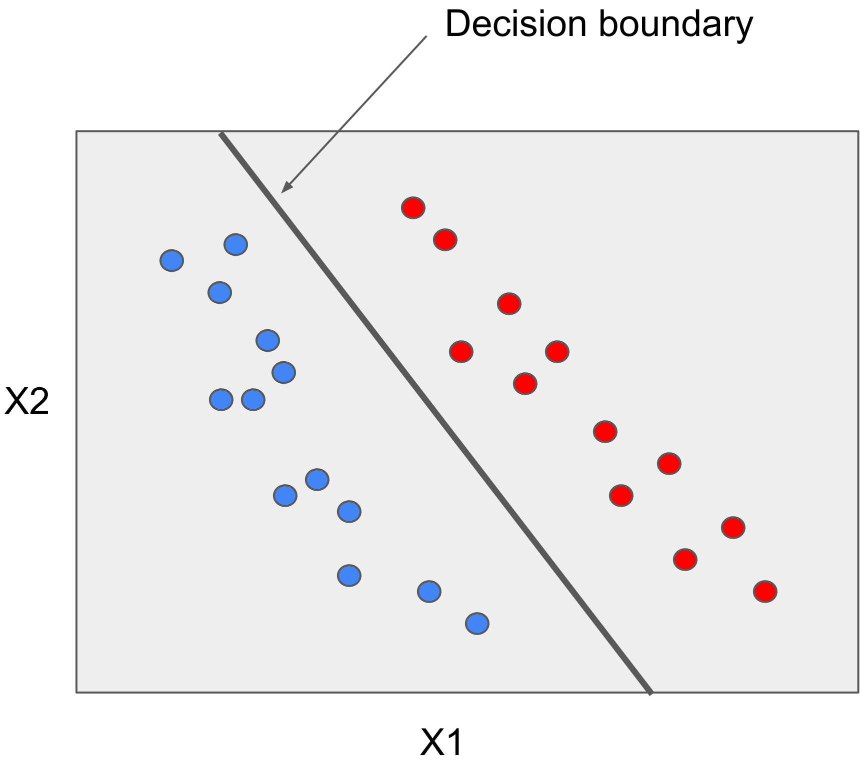
Explaining Local Decision Boundary



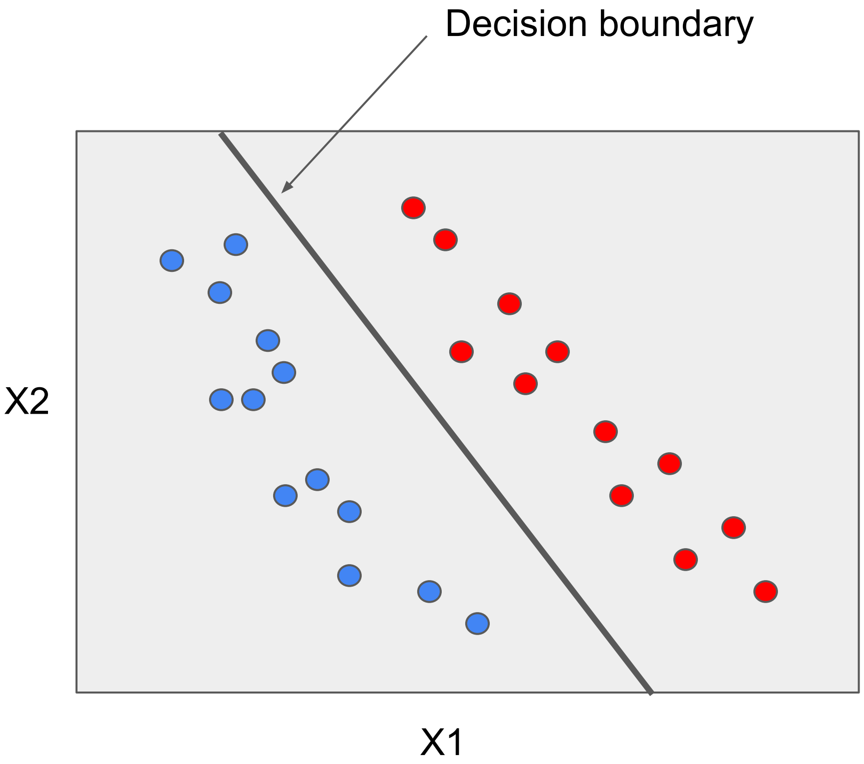
Explaining Local Decision Boundary



Explaining Local Decision Boundary



Explaining Local Decision Boundary



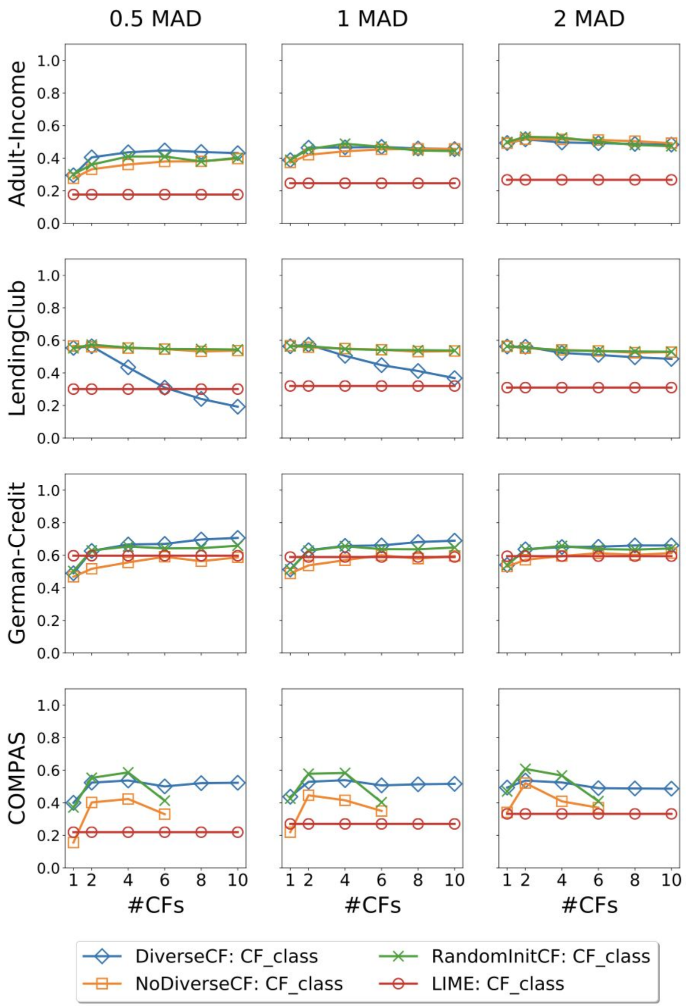
New goal: can we predict *f*'s outcomes using counterfactual and a simpler model (e.g., 1-NN)?

Approximating Decision Boundaries

- For different distances from original input
- Train models to predict *f*'s outcomes with discovered counterfactuals
  - DiverseCF: ours with 1-NN
  - NoDiverseCF: no diversity term with 1-NN
  - RandomInitCF
- Also compare with *LIME*
- Evaluate on F1 score

Approximating Decision Boundaries

- For different distances from original input
- Train models to predict *f*'s outcomes with discovered counterfactuals
  - DiverseCF: ours with 1-NN
  - NoDiverseCF: no diversity term with 1-NN
  - RandomInitCF
- Also compare with *LIME*
- Evaluate on F1 score
- Overall, DiCE performs better
  - Suggests it is better at finding local decision boundary



Causal Feasibility of CF Examples

- Potential counterfactual actions
  - Get married and get a master's degree and increase income by $50,000
  - Assumes age stays constant
- Actionable counterfactuals require time to make changes
- How can we design counterfactual generation engines to account for such causal dependencies between variables?
- Question for future research

Appendix

DPP Diversity

- AlexKulesza,BenTaskar,etal.2012.Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning 5, 2–3 (2012), 123–286.

Counterfactual Explanations Can Be Manipulated

Slack, Dylan, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. "Counterfactual explanations can be manipulated." *Advances in neural information processing systems* 34 (2021)

Outline

- Background
  - Counterfactual explanation
  - Recourse
  - Recourse fairness
- Overview of the paper
  - Key points
  - Setup
  - Objective and training
- Experiments and results
- Conclusions
- Appendix

Background

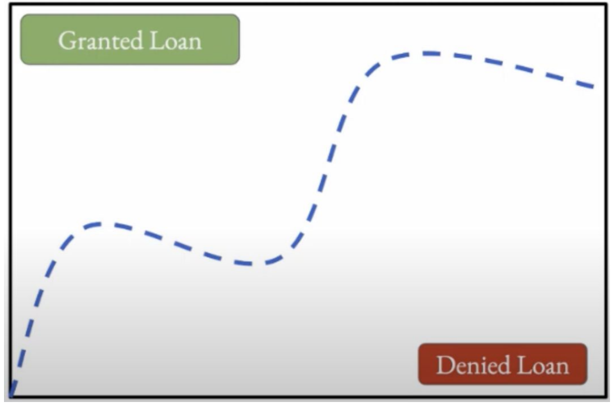
- Counterfactual Explanations:
    - A data point close to the original input
    - Predicted to be positive by the model

Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

- Objective in counterfactual algorithms:

Model f



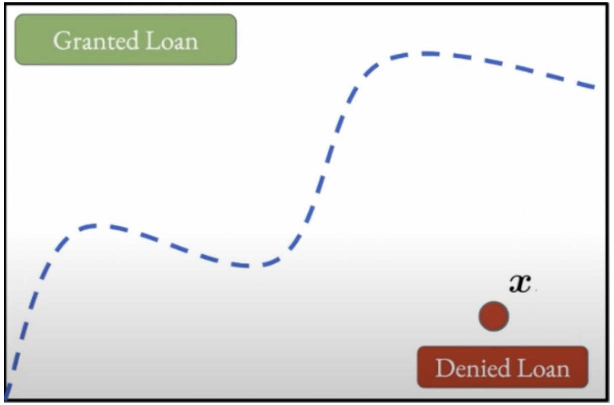
Source: slides of the paper at [slideslive]

Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

- Objective in counterfactual algorithms:

Model f



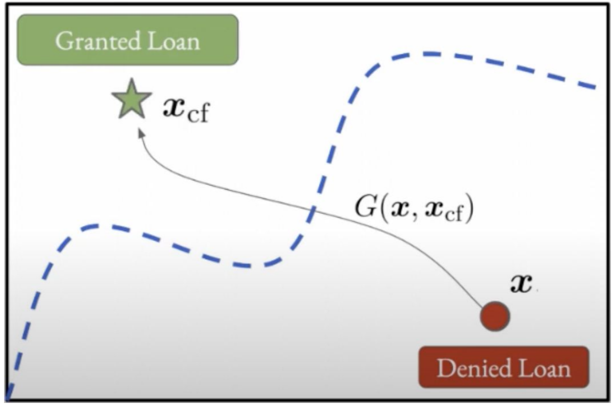
Source: slides of the paper at [slideslive]

Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

- Objective in counterfactual algorithms:

Model f



Source: slides of the paper at [slideslive]

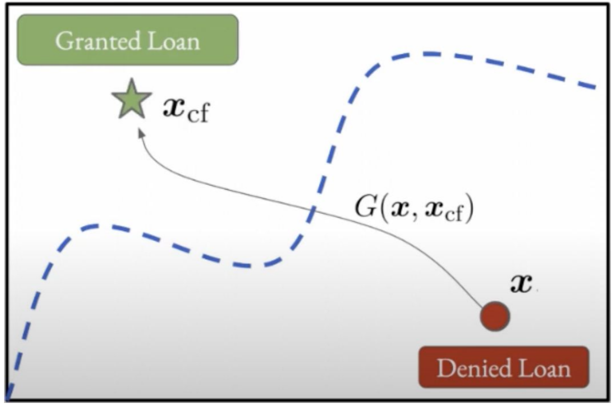
Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

- Objective in counterfactual algorithms:

$$G(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{cf}}) = \lambda \cdot (f(\boldsymbol{x}_{\mathrm{cf}}) - 1)^2 + d(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{cf}})$$

Model f



Source: slides of the paper at slideslive

Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

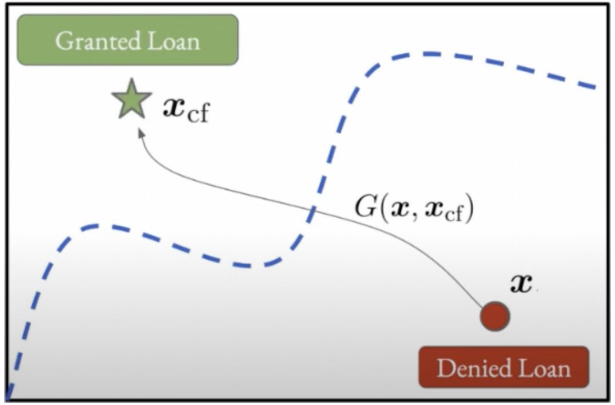
- Objective in counterfactual algorithms:

$$G(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{cf}}) = \boxed{\lambda \cdot (f(\boldsymbol{x}_{\mathrm{cf}}) - 1)^2} + \boxed{d(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{cf}})}$$

Encourages the desired outcome probability by the model

Encourages proximity

Model f



Source: slides of the paper at [slideslive]

Background

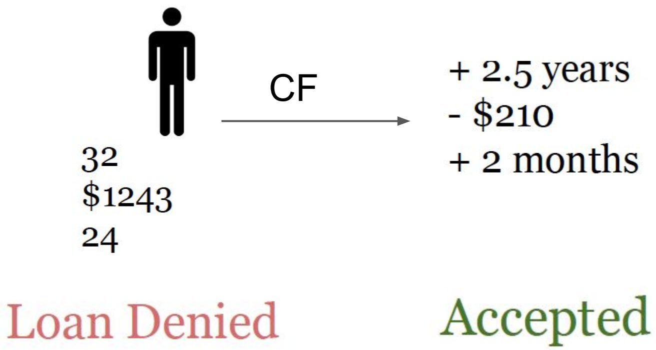
- Recourse: The difference between the original data point and the counterfactual

Background

- Recourse: The difference between the original data point and the counterfactual
- Example:
  - A 32 year-old male who wants to get a loan of $1243 for a duration of 24 months

Background

- Recourse: The difference between the original data point and the counterfactual
- Example:
  - A 32 year-old male who wants to get a loan of $1243 for a duration of 24 months
  - Counterfactual explanation: Had he been 2.5 years older and requested $210 less for a duration two months shorter, he would have been eligible for the loan.



Background

What if the counterfactual explanations return recourses that are easier to achieve for the *non-protected* group?

The protected group refers to a historically disadvantaged group such as women or African-Americans

Background

What if the counterfactual explanations return recourses that are easier to achieve for the non-protected group?

Unfairness in counterfactuals

How would you solve this problem?

Background

A model $f : x \rightarrow [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim \boxed{\mathcal{D}_{pr}^{neg}}} \left[\boxed{d} \left(\boldsymbol{x}, \boxed{\mathcal{A}(\boldsymbol{x})} \right) \right] - \mathbb{E}_{x \sim \boxed{\mathcal{D}_{np}^{neg}}} \left[d \left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x}) \right) \right] \right| \leq \tau$$

CF

Protected subset of the dataset with negative outcome

Distance function

Non-protected subset of the dataset with negative outcome

Background

A model $f : x \to [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim \mathcal{D}_{pr}^{neg}} \left[d \left(\boldsymbol{x}, \boxed{\mathcal{A}(\boldsymbol{x})} \right) \right] - \mathbb{E}_{x \sim \mathcal{D}_{np}^{neg}} \left[d \left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x}) \right) \right] \right| \leq \tau$$

CF

Protected subset of the dataset with negative outcome

Distance function

Non-protected subset of the dataset with negative outcome

Background

A model $f : x \to [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim \boxed{\mathcal{D}_{pr}^{neg}}} \left[\boxed{d} \left(\boldsymbol{x}, \boxed{\mathcal{A}(\boldsymbol{x})} \right) \right] - \mathbb{E}_{x \sim \boxed{\mathcal{D}_{np}^{neg}}} \left[d \left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x}) \right) \right] \right| \leq \tau$$

CF

Protected subset of the dataset with negative outcome

Distance function

Non-protected subset of the dataset with negative outcome

Background

A model $f : x \to [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim \boxed{\mathcal{D}_{pr}^{neg}}} \left[\boxed{d} \left(\boldsymbol{x}, \boxed{\mathcal{A}(\boldsymbol{x})}\right)\right] - \mathbb{E}_{x \sim \boxed{\mathcal{D}_{np}^{neg}}} \left[d \left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x})\right)\right] \right| \leq \tau$$

CF

Protected subset of the dataset with negative outcome

Distance function

Non-protected subset of the dataset with negative outcome

Background

Recourse fairness:

$$\left|\mathbb{E}_{x\sim\mathcal{D}_{pr}^{neg}}\left[d\left(\boldsymbol{x},\mathcal{A}(\boldsymbol{x})\right)\right]-\mathbb{E}_{x\sim\mathcal{D}_{np}^{neg}}\left[d\left(\boldsymbol{x},\mathcal{A}(\boldsymbol{x})\right)\right]\right|\leq\tau$$

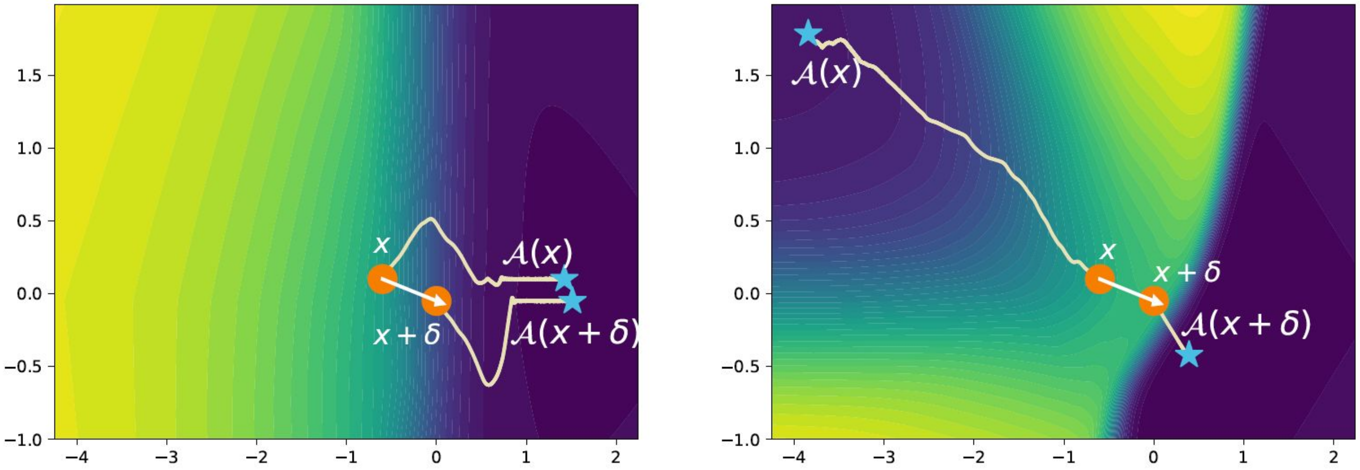
The costs of recourses for the protected and non-protected group should be close.

Key points of the paper

- Shows that counterfactual algorithms are not *robust.*
- Introduces a training objective for adversarial models.
- The adversarial models manipulate counterfactual explanations.

Key points of the paper

Counterfactual explanation search can converge to different local minima

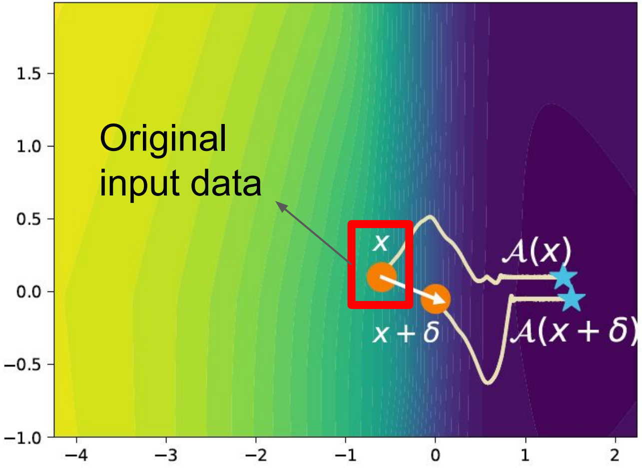


(a) **Training with BCE Objective**

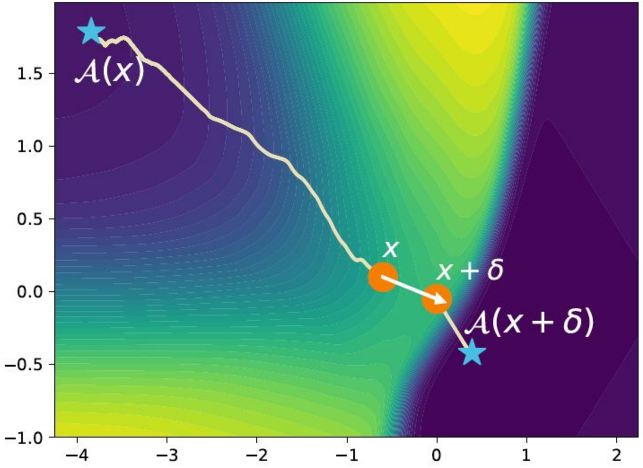
(b) **Training Adversarial Model**

Key points of the paper

Counterfactual explanation search can converge to different local minima



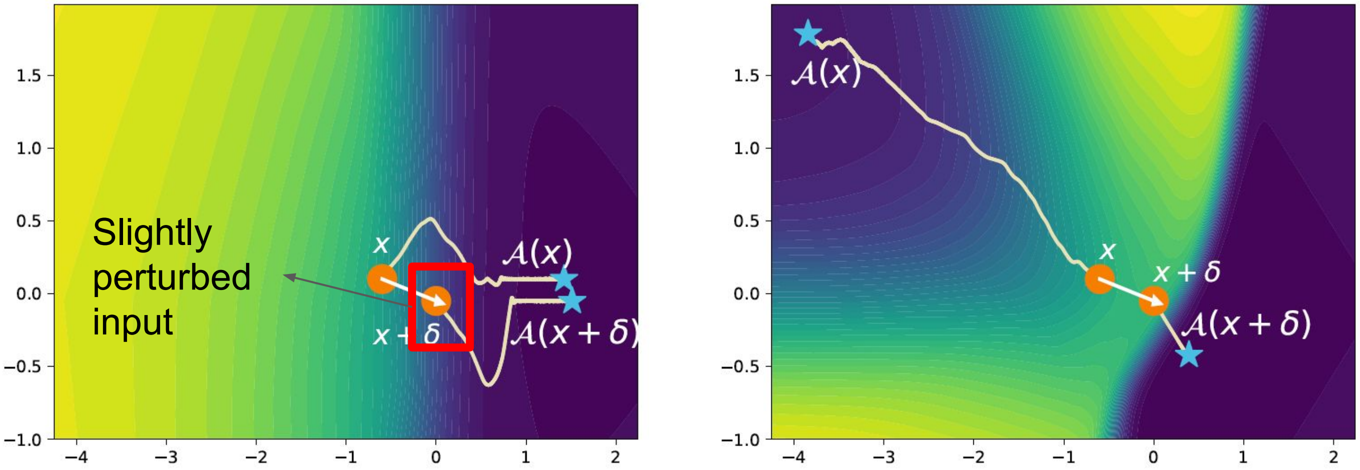
(a) **Training with BCE Objective**



(b) **Training Adversarial Model**

Key points of the paper

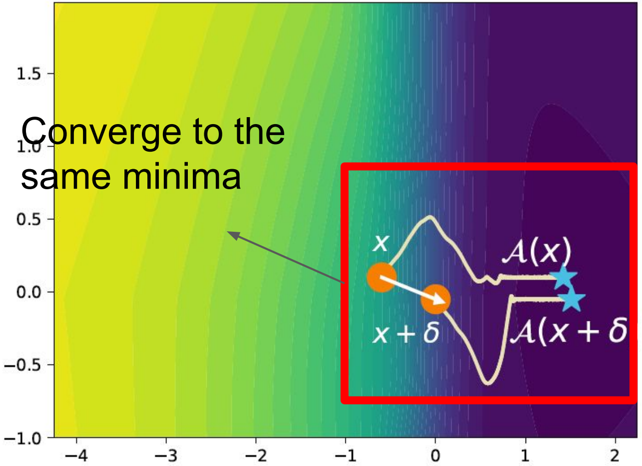
Counterfactual explanation search can converge to different local minima



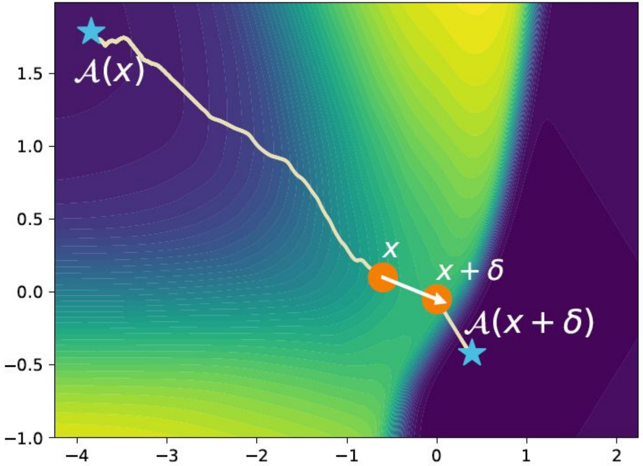
(a) **Training with BCE Objective** (b) **Training Adversarial Model**

Key points of the paper

Counterfactual explanation search can converge to different local minima



(a) **Training with BCE Objective**

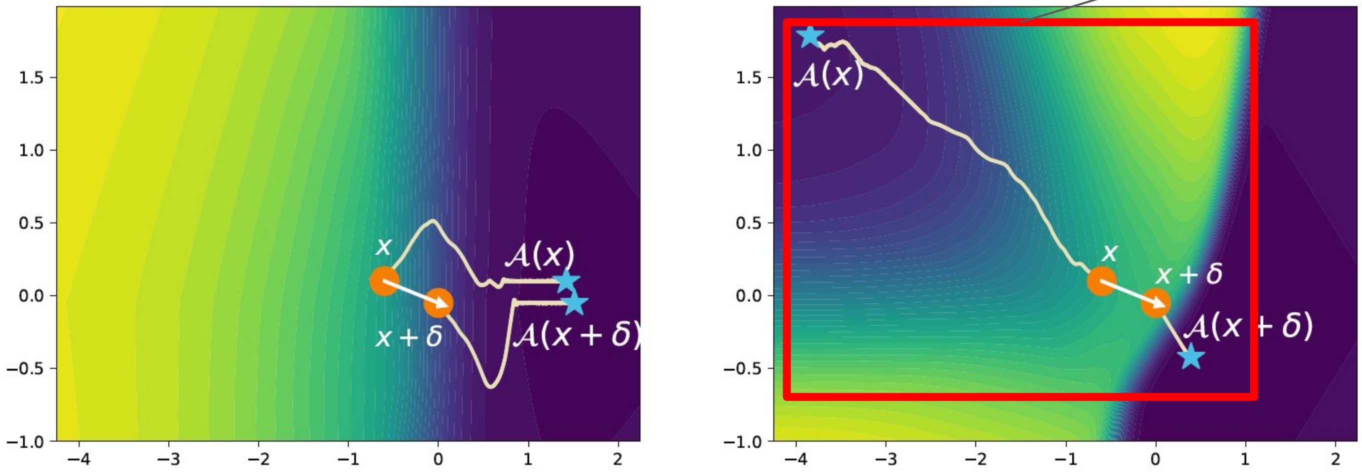


(b) **Training Adversarial Model**

Key points of the paper

Counterfactual explanation search can converge to different local minima

Recourse for the perturbed input is easier to achieve

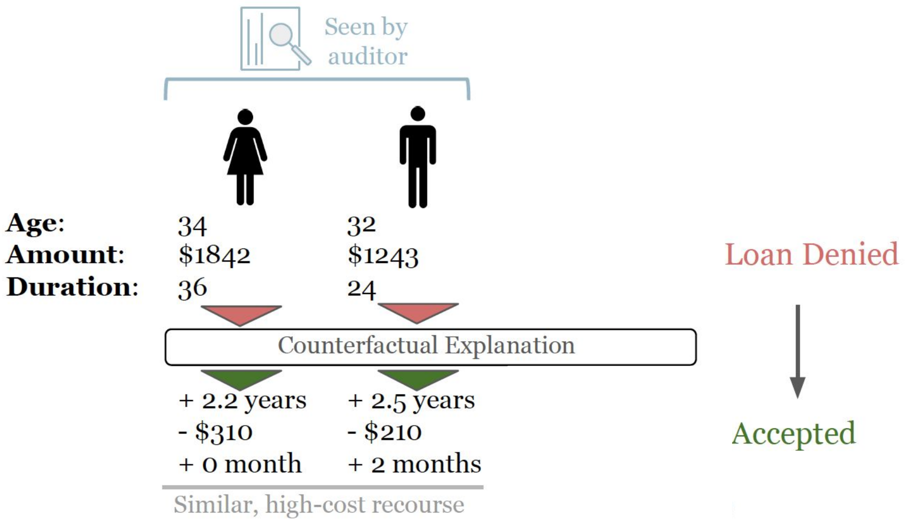


(a) **Training with BCE Objective** (b) **Training Adversarial Model**

How is this a vulnerability?

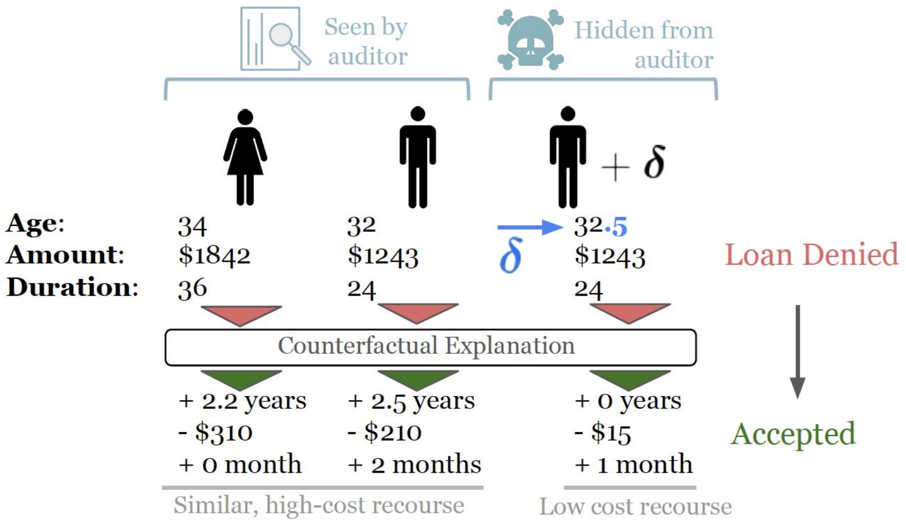
Counterfactual explanations can be manipulated

Example:



Counterfactual explanations can be manipulated

Example:



Setup

Adversarial model

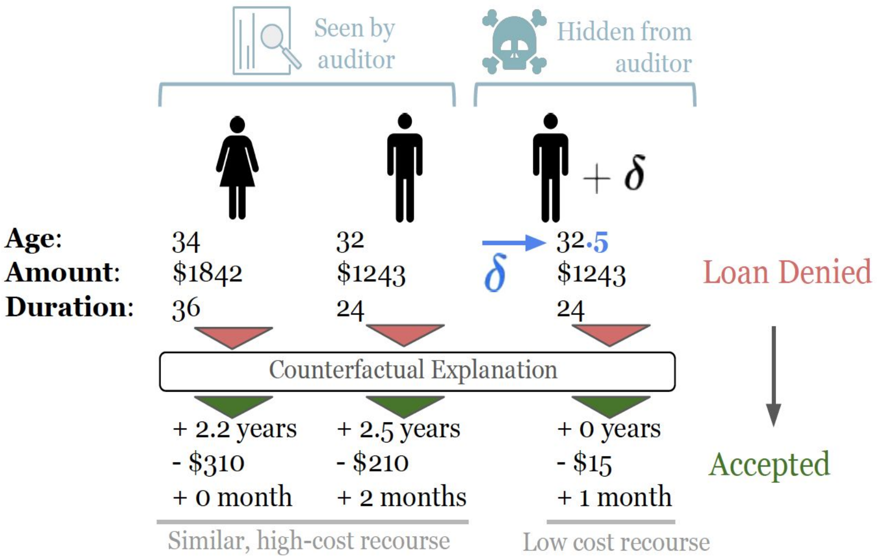
- Biased towards the non-protected group
- Passes the audits
- Produces very low cost counterfactuals for the non-protected group

Model auditor

- Makes sure the model is recourse fair

Training objective for adversarial model

- Fairness
- Unfairness
- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual



Training objective for adversarial model

- Fairness: Model should be fair according to this definition

$$\left| \mathbb{E}_{x \sim \mathcal{D}_{pr}^{neg}} \left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x})\right) \right] - \mathbb{E}_{x \sim \mathcal{D}_{np}^{neg}} \left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x})\right) \right] \right| \leq \tau$$

- Unfairness
- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual

Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{pr}}^{\text{neg}}}\left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x})\right)\right] \gg \mathbb{E}_{x \sim \mathcal{D}_{\text{np}}^{\text{neg}}}\left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x}+\boldsymbol{\delta})\right)\right]$$

- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual

Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse
- Small perturbation: Perturbation vectors should be small

$$\text{minimize}\ \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta})$$

- Accuracy
- Perturbed input should be a counterfactual

Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse
- Small perturbation: Perturbation vectors should be small
- Accuracy: Minimize the classification loss
- Perturbed input should be a counterfactual

Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse
- Small perturbation: Perturbation vectors should be small
- Accuracy: Minimize the classification loss
- Perturbed input should be a counterfactual

$$\text{minimize } \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x} + \boldsymbol{\delta}) - 1 \right)^2$$

Training the adversarial model

1. First stage:
    - Small perturbations
    - Counterfactuals under perturbations
    - Accuracy
    - Passes the perturbations and model weights to the second stage
2. Second stage:
    - Fairness
    - Unfairness
    - Accuracy

Training the adversarial model

1. First stage:

$$\boldsymbol{\delta} := \arg\min_{\boldsymbol{\delta}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x}+\boldsymbol{\delta}) - 1\right)^2 + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta})$$

Training the adversarial model

1. First stage:

$$\boldsymbol{\delta} := \arg\min_{\boldsymbol{\delta}} \min_{\boldsymbol{\theta}} \boxed{\mathcal{L}(\boldsymbol{\theta}, \mathcal{D})} + \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x} + \boldsymbol{\delta}) - 1\right)^2 + \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta})$$

Classification loss

Training the adversarial model

1. First stage:

$$\boldsymbol{\delta} := \arg\min_{\boldsymbol{\delta}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) + \boxed{\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x} + \boldsymbol{\delta}) - 1\right)^2} + \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta})$$

Perturbed input to be a counterfactual

Training the adversarial model

1. First stage:

$$\boldsymbol{\delta} := \arg\min_{\boldsymbol{\delta}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x}+\boldsymbol{\delta}) - 1\right)^2 + \boxed{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta})}$$

Perturbation should be small

Training the adversarial model

1. First stage:

$$\boldsymbol{\delta} := \arg\min_{\boldsymbol{\delta}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left(f(\boldsymbol{x}+\boldsymbol{\delta}) - 1\right)^2 + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} d(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta})$$

2. Second stage:

$$\boldsymbol{\theta} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left[d\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta})\right)\right] + \left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{pr}}^{\text{neg}}} \left[d\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left[d\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]\right)^2$$

$$\text{s.t.} \quad \mathbb{E}_{x\sim\mathcal{D}_{\text{np}}^{\text{neg}}} \left[d\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta})\right)\right] < \mathbb{E}_{x\sim\mathcal{D}_{\text{pr}}^{\text{neg}}} \left[d\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

Experiments

- Dataset
  - Used two datasets: “German Credit” and “Communities and Crimes”
  - Strong incentives to “game the system” in both datasets

Experiments

- Dataset
  - Used two datasets: “German Credit” and “Communities and Crimes”
  - Strong incentives to “game the system” in both datasets
- Manipulated Model
  - 4 layer feed-forward neural network
  - Tanh activation function
  - Adam optimizer and cross entropy loss

How do you expect the accuracy to be impacted in the manipulated model?

Results

Impact of the manipulated model on accuracy:

	Comm. & Crime		**German Credit**	
	Acc	$\|\|\boldsymbol{\delta}\|\|_1$	Acc	$\|\|\boldsymbol{\delta}\|\|_1$
Unmodified	81.2	-	71.1	-
Wachter et al.	80.9	0.80	72.0	0.09
Sparse Wachter	77.9	0.46	70.5	2.50
Prototypes	79.2	0.46	69.0	2.21
DiCE	81.1	1.73	71.2	0.09

Results

- Metrics
  - Effectiveness of manipulation

$$\text{Cost reduction} := \frac{\mathbb{E}_{x\sim\mathcal{D}_{\text{np}}^{\text{neg}}}\left[d(\boldsymbol{x},\mathcal{A}(\boldsymbol{x}))\right]}{\mathbb{E}_{x\sim\mathcal{D}_{\text{np}}^{\text{neg}}}\left[d(\boldsymbol{x},\mathcal{A}(\boldsymbol{x}+\boldsymbol{\delta}))\right]}$$

Results

- Metrics
    - Effectiveness of manipulation

Table 2: **Recourse Costs of Manipulated Models**: Counterfactual algorithms find similar cost recourses for both subgroups, however, give much lower cost recourse if δ is added before the search.

	Communities and Crime				German Credit			
	Wach.	**S-Wach.**	**Proto.**	**DiCE**	**Wach.**	**S-Wach.**	**Proto.**	**DiCE**
Protected	35.68	54.16	22.35	49.62	5.65	8.35	10.51	6.31
Non-Protected	35.31	52.05	22.65	42.63	5.08	8.59	13.98	6.81
Disparity	*0.37*	*2.12*	*0.30*	*6.99*	*0.75*	*0.24*	*0.06*	*0.5*
Non-Protected+δ	1.76	22.59	8.50	9.57	3.16	4.12	4.69	3.38
Cost reduction	*20.1×*	*2.3×*	*2.6×*	*4.5×*	*1.8×*	*2.0×*	*2.2×*	*2.0×*

Results

- Metrics
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?

$$P(\mathcal{A}(\boldsymbol{x})) = \frac{d(\mathcal{A}(\boldsymbol{x}), a_0)}{\min_{\boldsymbol{x} \neq a_0 \in \mathcal{D}_{\text{pos}} \cap \{\forall x \in \mathcal{D}_{\text{pos}} \mid f(x)=1\}} d(a_0, \boldsymbol{x})}$$

Results

- Metrics
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?

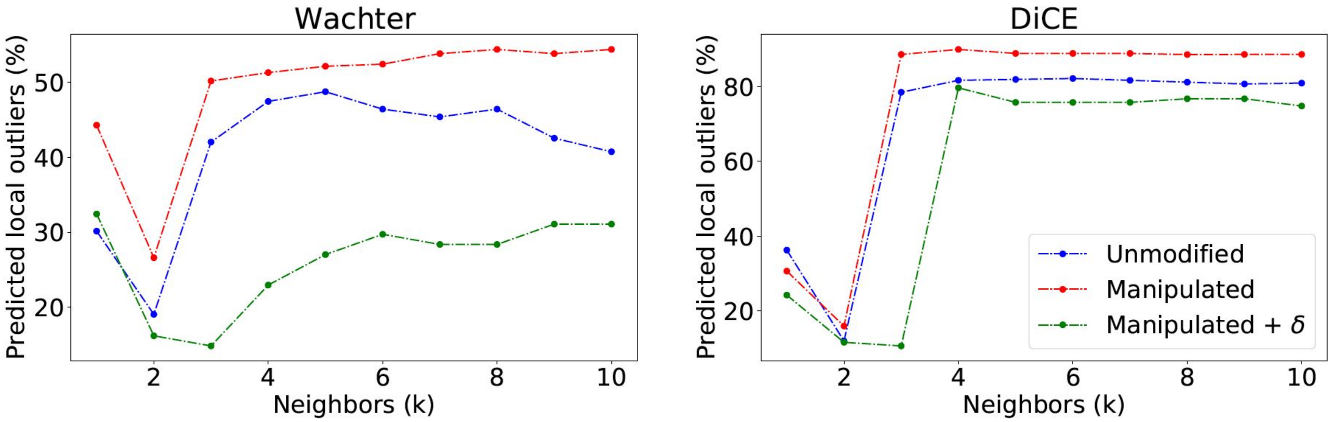
$$P(\mathcal{A}(\boldsymbol{x})) = \frac{d(\mathcal{A}(\boldsymbol{x}), a_0)}{\min_{\boldsymbol{x} \neq a_0 \in \mathcal{D}_{\text{pos}} \cap \{\forall x \in \mathcal{D}_{\text{pos}} | f(x)=1\}} d(a_0, \boldsymbol{x})}$$

The local outlier factor of the counterfactuals with respect to the positively classified data (Breunig et al. “LOF: identifying density-based local outliers”)

Will be >1 if the counterfactual is an outlier.

Results

- Metrics
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?



Conclusions

- The paper shows that counterfactual explanations can be manipulated.
- They train an adversarial model that produces seemingly fair recourses but is in fact biased towards the non-protected group.
- They show that the manipulations are effective and realistic.

Appendix

How to train the adversarial model if the counterfactual algorithm is black box?

Lemma 3.1 *Assuming the counterfactual explanation $\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})$ follows the form of the objective in Equation 1, $\frac{\partial}{\partial \boldsymbol{x}_{cf}} G(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})) = 0$, and m is the number of parameters in the model, we can write the derivative of counterfactual explanation $\mathcal{A}$ with respect to model parameters $\boldsymbol{\theta}$ as the Jacobian,*

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\left[\frac{\partial^2 G\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)}{d\boldsymbol{x}_{cf}^2}\right]^{-1} \cdot \left[\frac{\partial}{\partial \boldsymbol{\theta}_1} \frac{\partial}{\partial \boldsymbol{x}_{cf}} G\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \cdots \frac{\partial}{\partial \boldsymbol{\theta}_m} \frac{\partial}{\partial \boldsymbol{x}_{cf}} G\left(\boldsymbol{x}, \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$