

Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

Ramaravind Kommiya Mothilal

Microsoft Research India
t-rakom@microsoft.com

Amit Sharma

Microsoft Research India
amshar@microsoft.com

Chenhao Tan

University of Colorado Boulder
chenhao.tan@colorado.edu



Explaining machine learning predictions

Techniques to explain machine predictions

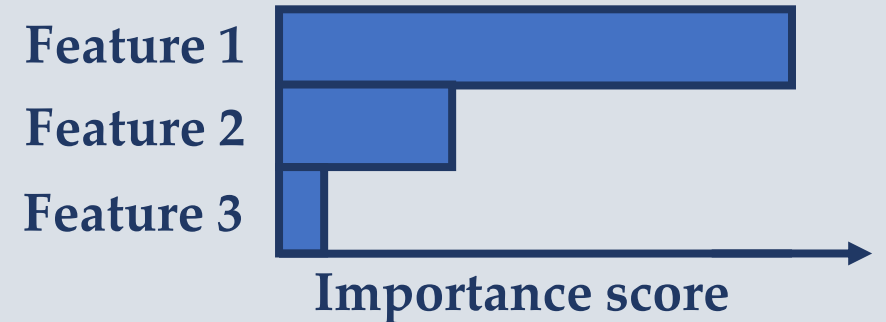
LIME (Ribeiro et al., 2016); **Local Rule-based** (Guidotti et al., 2018);
SHAP (Lundberg et al., 2017); **Intelligible Models** (Lou et al., 2012);

Explaining machine learning predictions

Techniques to explain machine predictions

LIME (Ribeiro et al., 2016); **Local Rule-based** (Guidotti et al., 2018);
SHAP (Lundberg et al., 2017); **Intelligible Models** (Lou et al., 2012);

Feature importance-based methods are widely used in many practical applications



But there is an important problem...

But, what a decision-subject should do to get a **desired** outcome?



But there is an important problem...

But, what a decision-subject should do to get a **desired** outcome?

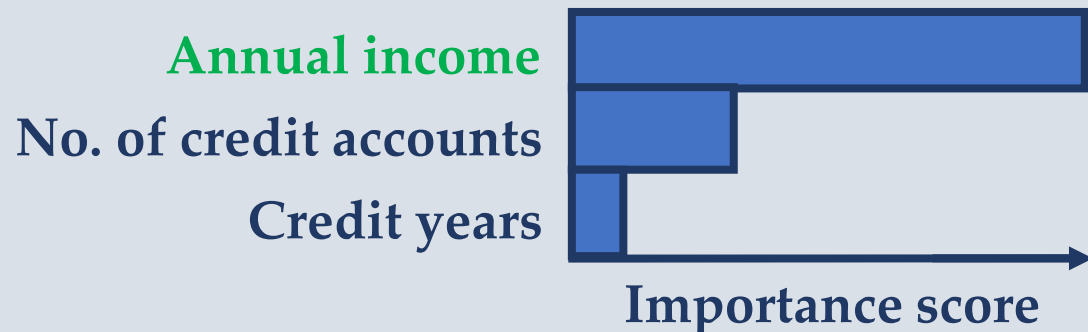


But there is an important problem...

But, what a decision-subject should do to get a **desired** outcome?

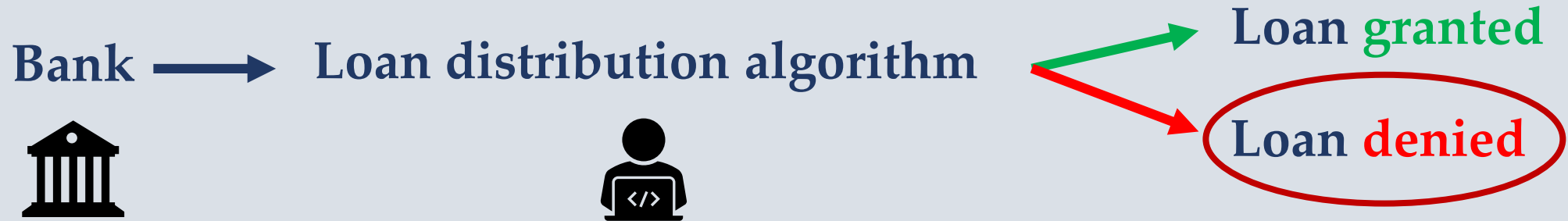


Feature importance-based explanations

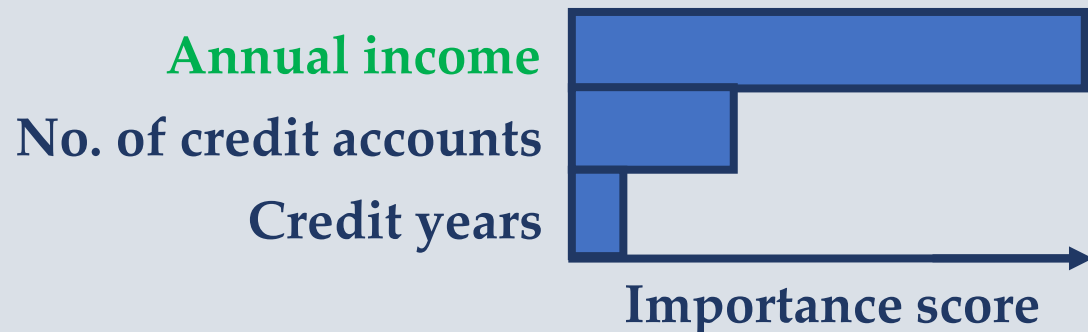


But there is an important problem...

But, what a decision-subject should do to get a **desired** outcome?



Feature importance-based explanations



Counterfactual explanations (CF)

("what-if" scenarios) (Wachter et al., 2017)

You would have got the loan if your **annual income had been 100,000**

Desirable properties for counterfactuals

Actionability :

Users should be able to make the changes indicated by counterfactuals

Desirable properties for counterfactuals

Actionability :

Users should be able to make the changes indicated by counterfactuals

Feasibility

- ✓ Proximity
- ✓ User constraints
- ✓ Sparsity
- ✓ Causal constraints

Desirable properties for counterfactuals

Actionability :

Users should be able to make the changes indicated by counterfactuals

Feasibility

+

Diversity

- ✓ Proximity
- ✓ User constraints
- ✓ Sparsity
- ✓ Causal constraints

Desirable properties for counterfactuals

Actionability :

Users should be able to make the changes indicated by counterfactuals

Feasibility

+

Diversity

- ✓ Proximity
- ✓ User constraints
- ✓ Sparsity
- ✓ Causal constraints

Wachter et al (2017)

$$C = \arg \min_c \mathbf{yloss}(f(c), \mathbf{y}) + |\mathbf{x} - c|$$

Russell (2017)

Mixed integer programming

Works only for linear ML models

General optimization framework

Diverse
counterfactual
explanations



$$\mathbf{C}(x) = \arg \min_{c_1, \dots, c_k} \frac{1}{k} \sum_{i=1}^k \mathbf{yloss}(f(c_i), y) + \frac{\lambda_1}{k} \mathbf{dist}(c_i, x) - \lambda_2 \mathbf{dpp_diversity}(c_1, \dots, c_k)$$

k – no. of counterfactuals

λ_1 and λ_2 – loss-balancing hyperparameters

Loss to get
**desirable
outcome**



Loss to ensure
proximity to
original input



Loss to provide
diverse
explanations



$$\mathbf{dpp_diversity} = \det(\mathbf{K}),$$

$$\mathbf{K} = \frac{1}{1 + \mathbf{dist}(\mathbf{c}_i, \mathbf{c}_j)}$$

Practical considerations

$$\mathbf{C}(\mathbf{x}) = \arg \min_{c_1, \dots, c_k} \frac{1}{k} \sum_{i=1}^k \mathbf{yloss}(f(c_i), y) + \frac{\lambda_1}{k} \mathbf{dist}(c_i, \mathbf{x}) - \lambda_2 \mathbf{dpp_diversity}(c_1, \dots, c_k)$$

Practical considerations

$$\mathbf{C}(\mathbf{x}) = \arg \min_{c_1, \dots, c_k} \frac{1}{k} \sum_{i=1}^k \mathbf{yloss}(f(c_i), y) + \frac{\lambda_1}{k} \mathbf{dist}(c_i, x) - \lambda_2 \mathbf{dpp_diversity}(c_1, \dots, c_k)$$

- ❑ Incorporate additional feasibility properties
 - a) **Sparsity** – post-hoc correction
 - b) **User constraints**
- ❑ Choice of yloss – **hinge** loss
- ❑ Separate categorical and continuous distance functions
- ❑ Relative scale of mixed features

Practical considerations

$$\mathbf{C}(\mathbf{x}) = \arg \min_{c_1, \dots, c_k} \frac{1}{k} \sum_{i=1}^k \mathbf{yloss}(f(c_i), y) + \frac{\lambda_1}{k} \mathbf{dist}(c_i, \mathbf{x}) - \lambda_2 \mathbf{dpp_diversity}(c_1, \dots, c_k)$$

- ❑ Incorporate additional feasibility properties
 - a) **Sparsity** – post-hoc correction
 - b) **User constraints**
- ❑ Choice of yloss – **hinge** loss
- ❑ Separate categorical and continuous distance functions
- ❑ Relative scale of mixed features

Python library

DiCE

(**D**iverse **C**ounterfactual **E**xplanations)

<https://github.com/microsoft/DiCE>

Diverse counterfactual explanations

Adult-Income:

Predicting income based on demographical and educational variables
(UCI ML repository)

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	—

Quantitative evaluation framework for any counterfactual method

Quantitative evaluation framework for any counterfactual method

➤ **Metrics for comparison**

Quantitative evaluation framework for any counterfactual method

➤ Metrics for comparison

➤ Approximate local decision boundary

Quantitative evaluation framework for any counterfactual method

➤ Metrics for comparison

□ Validity : $\frac{|\{\text{unique instances in } C \text{ s.t. } f(c) > 0.5\}|}{k}$

□ Proximity : $1 - \frac{1}{k} \sum_{i=1}^k \text{dist}_{\text{cat}}(c_i, x)$

□ Sparsity : $1 - \frac{1}{kd} \sum_{i=1}^k \sum_{l=1}^d 1[c_i^l \neq x_i^l]$

□ Diversity : $\frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i}^k \sum_{l=1}^d 1[c_i^l \neq c_j^l]$

➤ Approximate local decision boundary

Evaluation metrics are not the same
as CF generation metrics

Quantitative evaluation framework for any counterfactual method

➤ Metrics for comparison

□ Validity : $\frac{|\{\text{unique instances in } C \text{ s.t. } f(c) > 0.5\}|}{k}$

□ Proximity : $1 - \frac{1}{k} \sum_{i=1}^k \text{dist}_{\text{cat}}(c_i, x)$

□ Sparsity : $1 - \frac{1}{kd} \sum_{i=1}^k \sum_{l=1}^d 1[c_i^l \neq x_i^l]$

□ Diversity : $\frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i}^k \sum_{l=1}^d 1[c_i^l \neq c_j^l]$

➤ Approximate local decision boundary

Can CF explanations help users “**extrapolate**” the local decision boundary of the ML model?

Evaluation metrics are not the same
as CF generation metrics

Quantitative evaluation framework for any counterfactual method

➤ Metrics for comparison

□ Validity : $\frac{|\{\text{unique instances in } C \text{ s.t. } f(c) > 0.5\}|}{k}$

□ Proximity : $1 - \frac{1}{k} \sum_{i=1}^k \text{dist}_{\text{cat}}(c_i, x)$

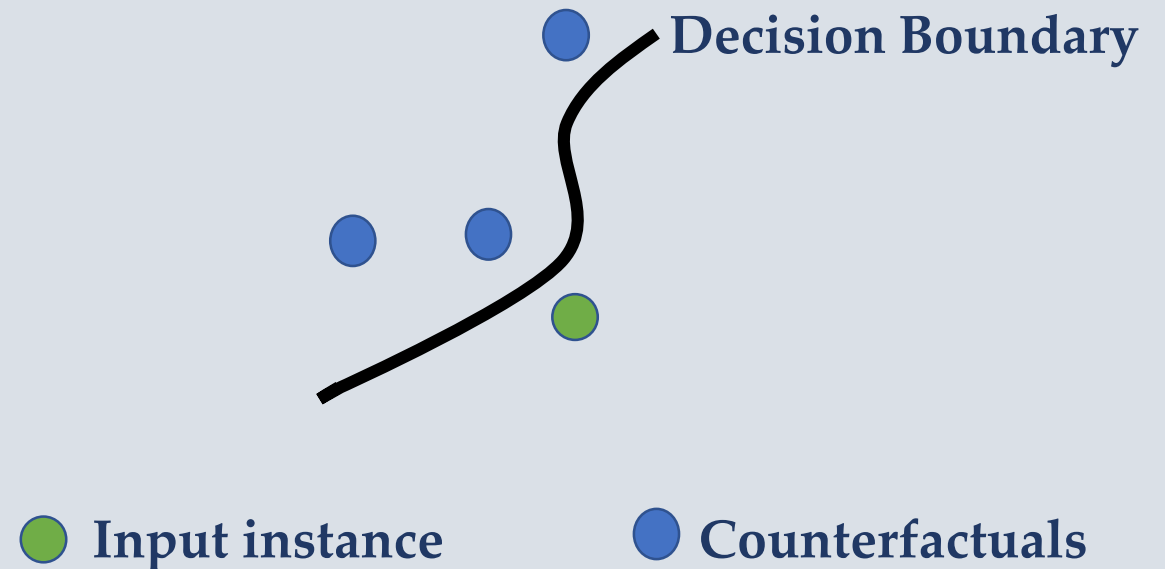
□ Sparsity : $1 - \frac{1}{kd} \sum_{i=1}^k \sum_{l=1}^d 1[c_i^l \neq x_i^l]$

□ Diversity : $\frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i}^k \sum_{l=1}^d 1[c_i^l \neq c_j^l]$

Evaluation metrics are not the same as CF generation metrics

➤ Approximate local decision boundary

Can CF explanations help users “**extrapolate**” the local decision boundary of the ML model?



Quantitative evaluation framework for any counterfactual method

➤ Metrics for comparison

□ Validity : $\frac{|\{\text{unique instances in } C \text{ s.t. } f(c) > 0.5\}|}{k}$

□ Proximity : $1 - \frac{1}{k} \sum_{i=1}^k \text{dist}_{\text{cat}}(c_i, x)$

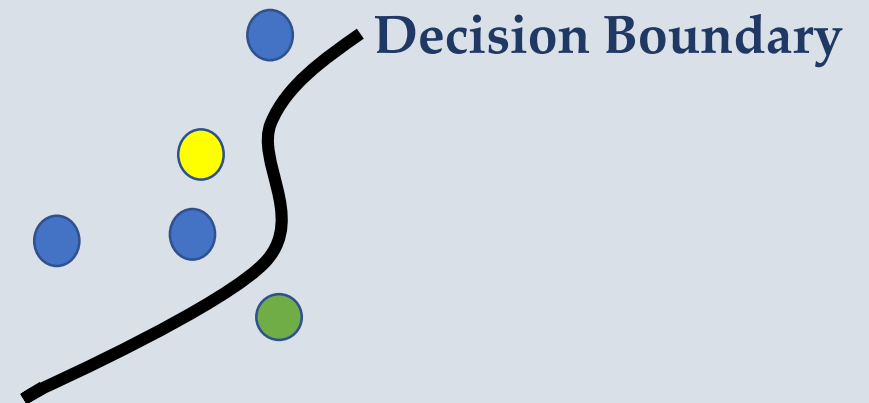
□ Sparsity : $1 - \frac{1}{kd} \sum_{i=1}^k \sum_{l=1}^d 1[c_i^l \neq x_i^l]$

□ Diversity : $\frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i}^k \sum_{l=1}^d 1[c_i^l \neq c_j^l]$

Evaluation metrics are not the same as CF generation metrics

➤ Approximate local decision boundary

Can CF explanations help users “**extrapolate**” the local decision boundary of the ML model?



● Input instance

● Counterfactuals

● Test instance

Results: comparing CF-based methods

Results: comparing CF-based methods

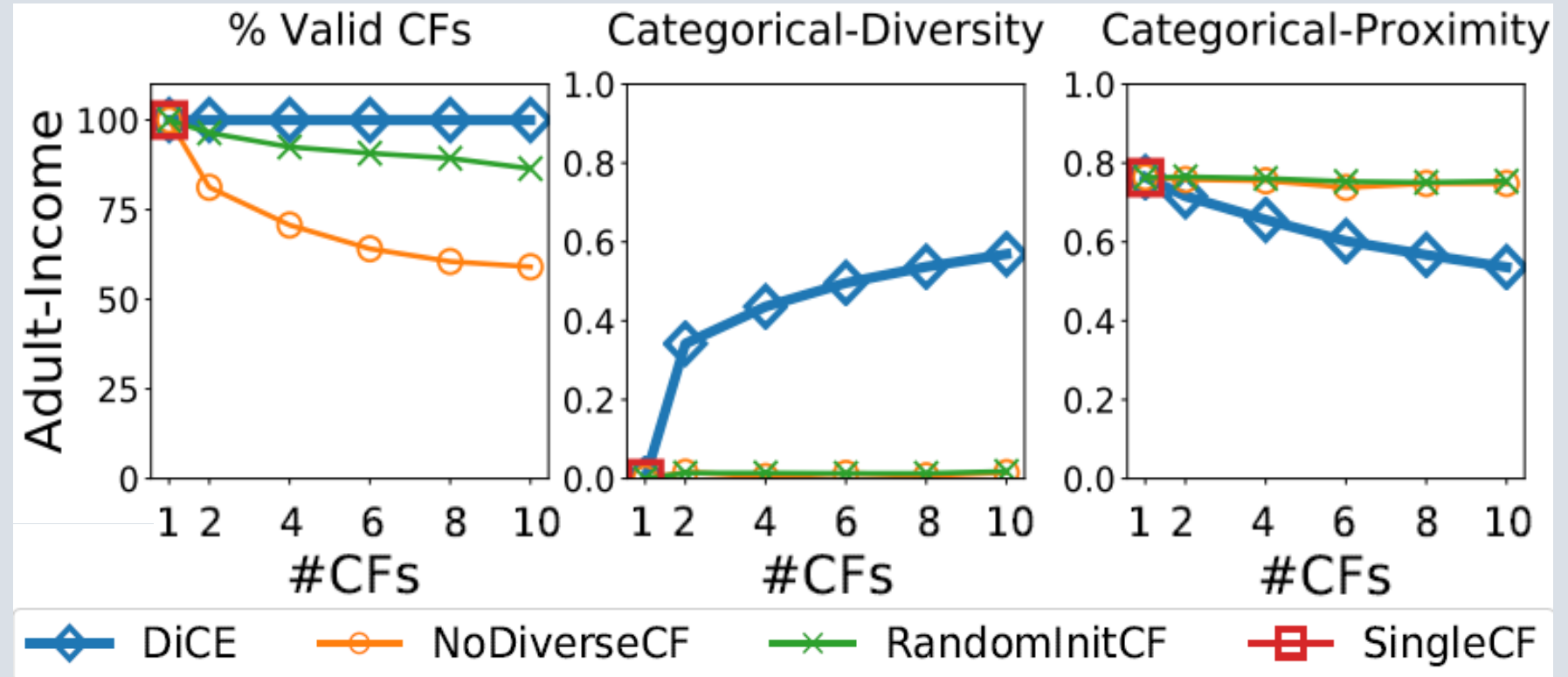
Explaining nonlinear ML models [compared to Wachter et al. (2017) and baselines]

Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit



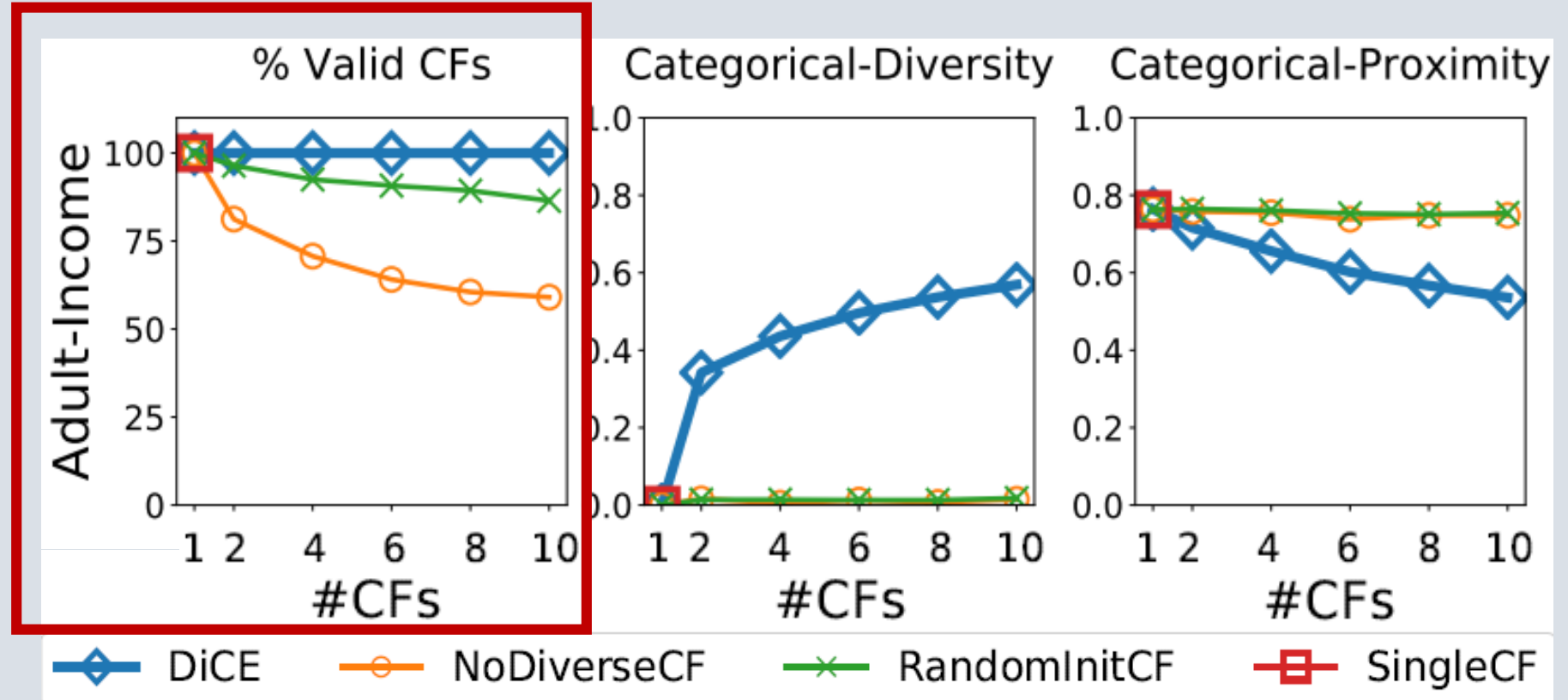
Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit

- 100% valid CFs till k=10



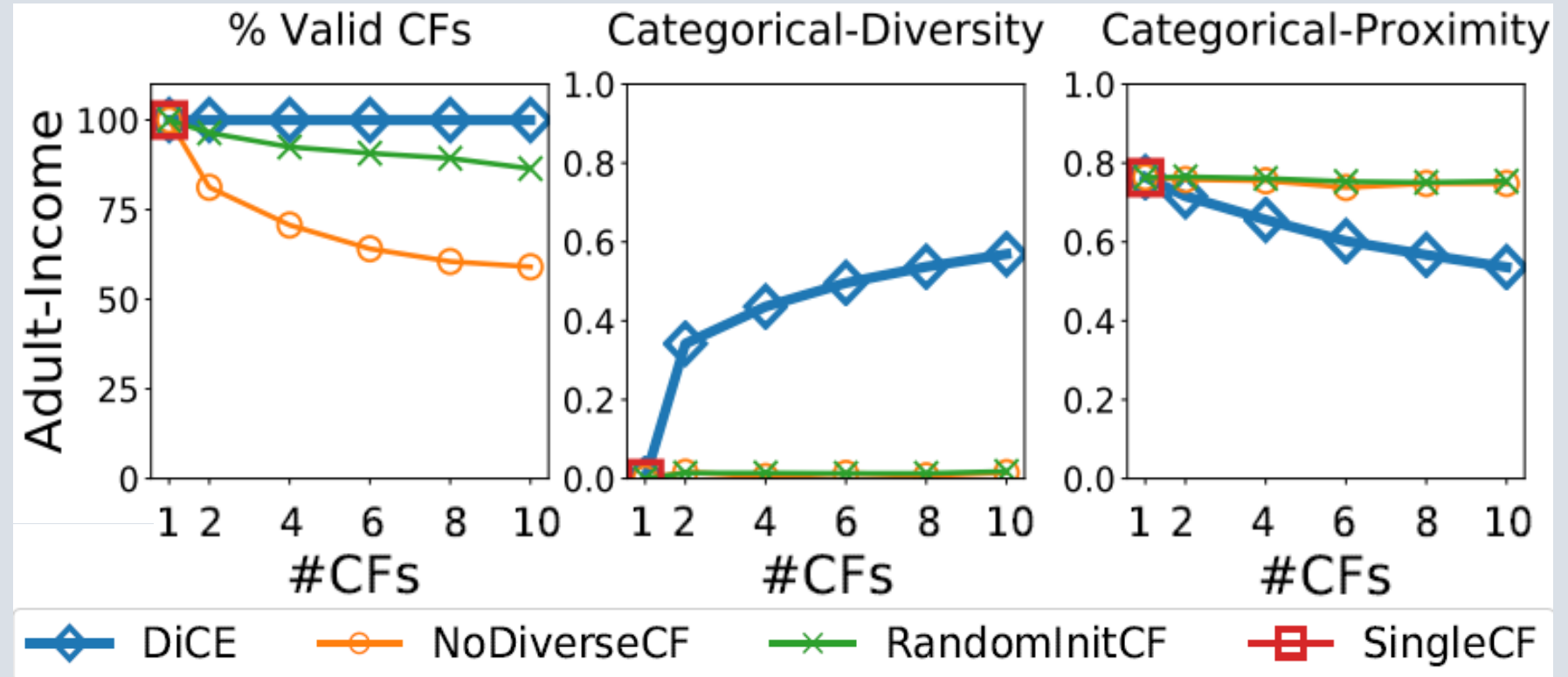
Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit

- 100% valid CFs till k=10

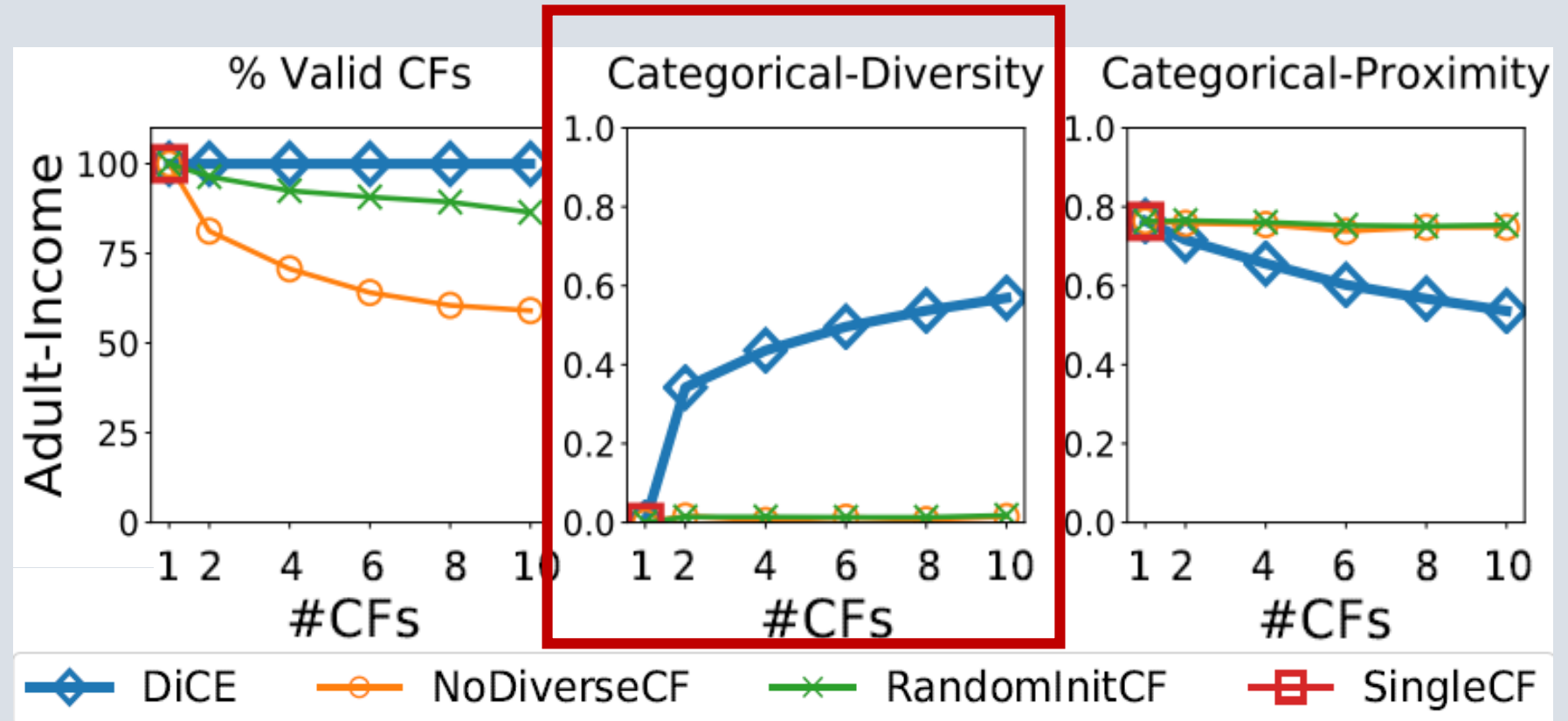


Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
 - COMPAS
 - Lending-Club
 - German-Credit
- 100% valid CFs till k=10
- Higher diversity

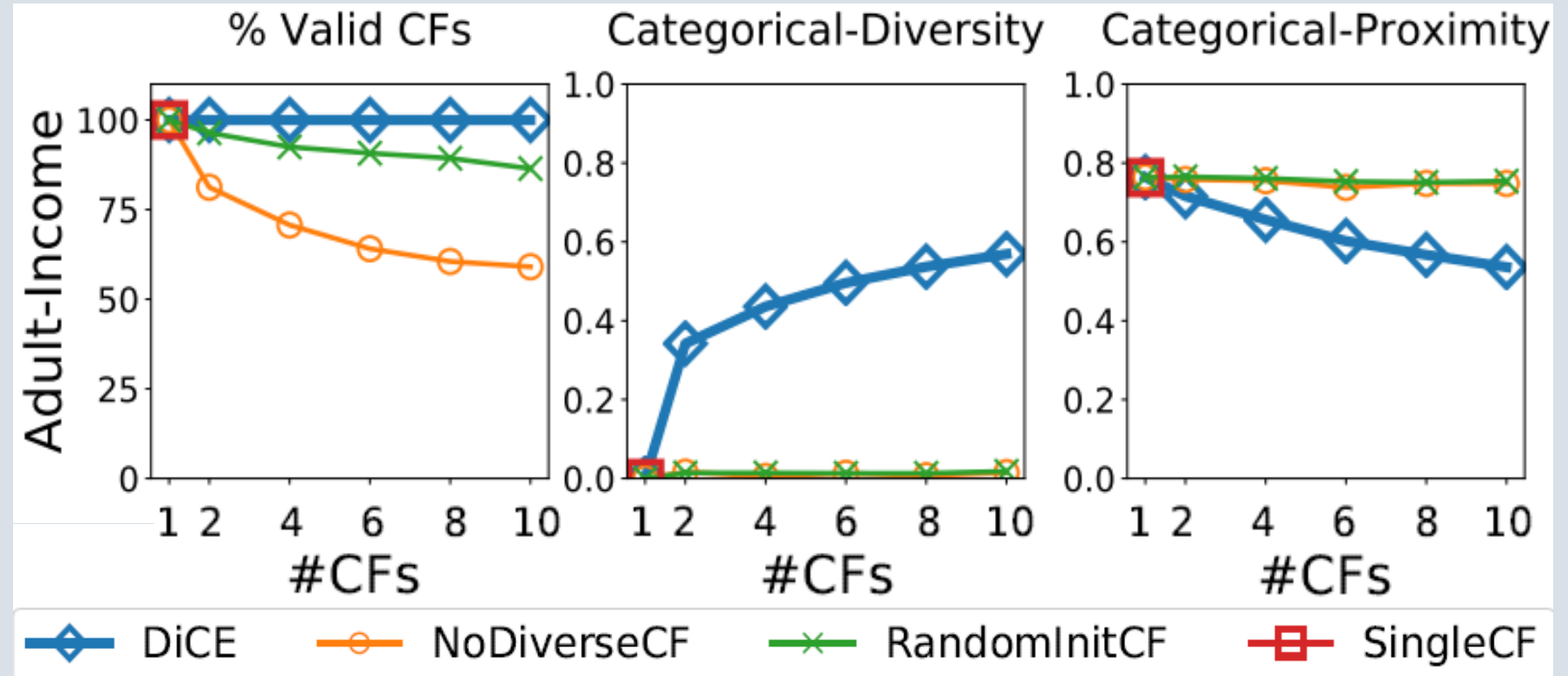


Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
 - COMPAS
 - Lending-Club
 - German-Credit
- 100% valid CFs till k=10
- Higher diversity



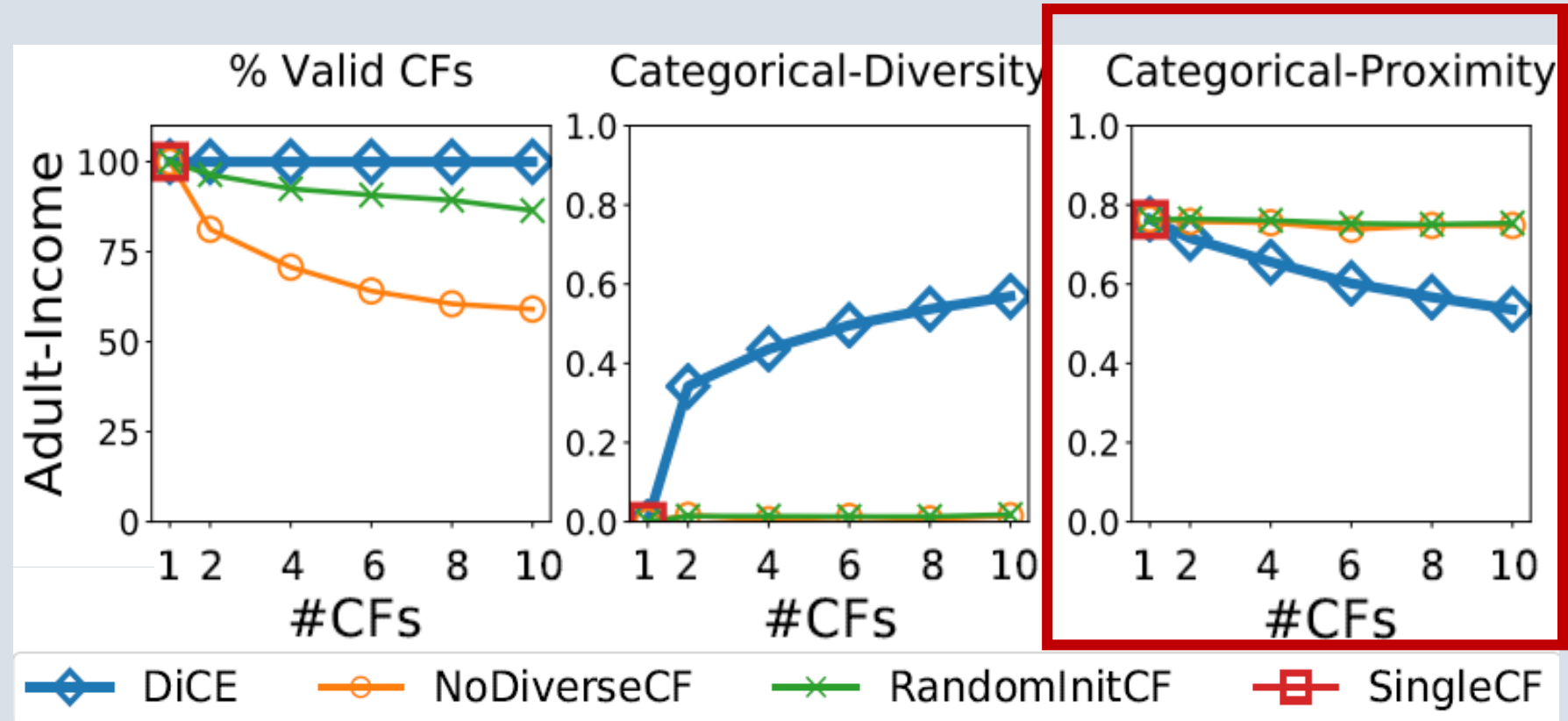
Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit

- 100% valid CFs till k=10
- Higher diversity
- Lower proximity



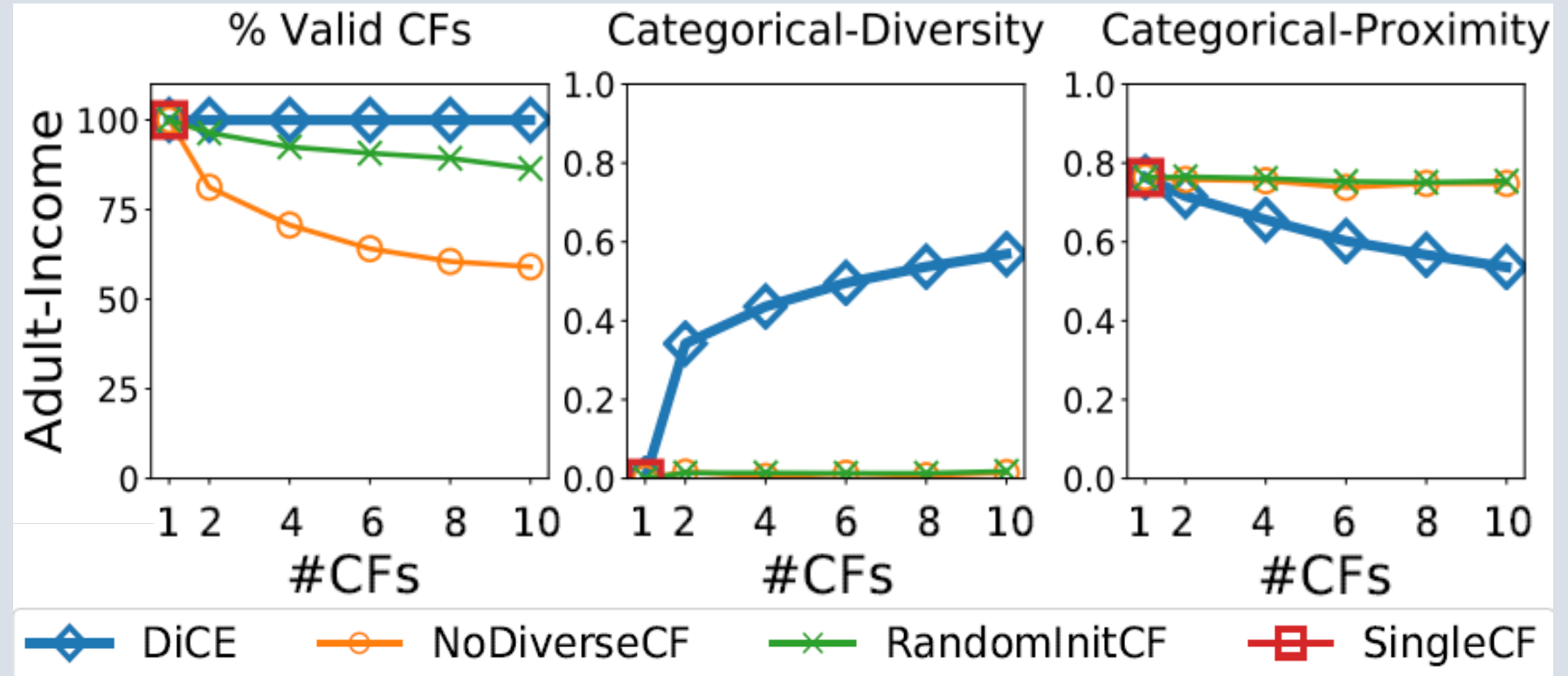
Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit

- 100% valid CFs till k=10
- Higher diversity
- Lower proximity



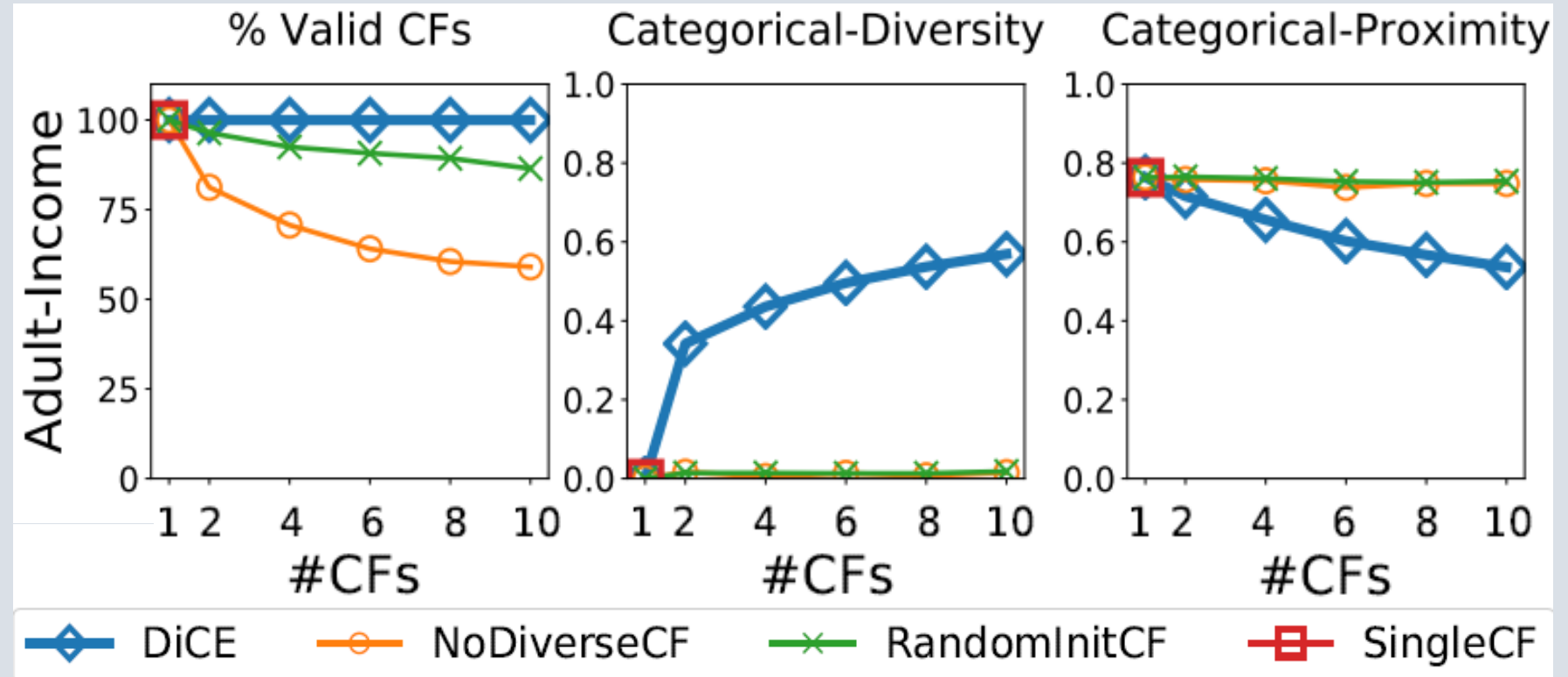
Results: comparing CF-based methods

Explaining nonlinear ML models [compared to [Wachter et al. \(2017\)](#) and baselines]

Datasets:

- [Adult-Income](#)
- COMPAS
- Lending-Club
- German-Credit

- 100% valid CFs till k=10
- Higher diversity
- Lower proximity

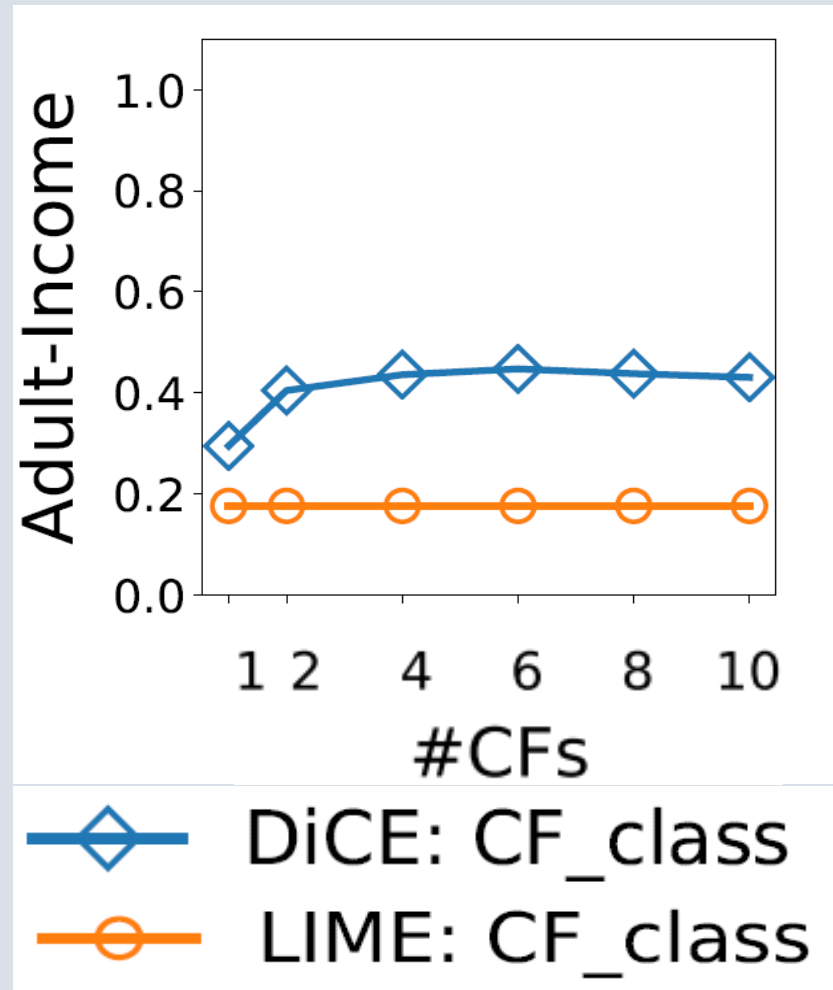


Explaining linear ML models [compared to [Russel \(2019\)](#) and baselines]

- More diversity and proximity

Results: approximating local decision boundary

F1 score



- Our method can approximate the local decision boundary **at least as well as** local explanation methods like LIME
- With **≤ 10 CFs** while LIME is based on **5000 samples**.

Summary and Future Work

Our Contributions

- **Diverse Counterfactual Explanations**
<https://github.com/microsoft/DiCE>
- **Quantitative** evaluation framework

Future Work

- Support for **fully black-box ML** models
- Incorporate **causal knowledge** during CF generation
- Make it useful for **different stakeholders** of explanations