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

Microsoft

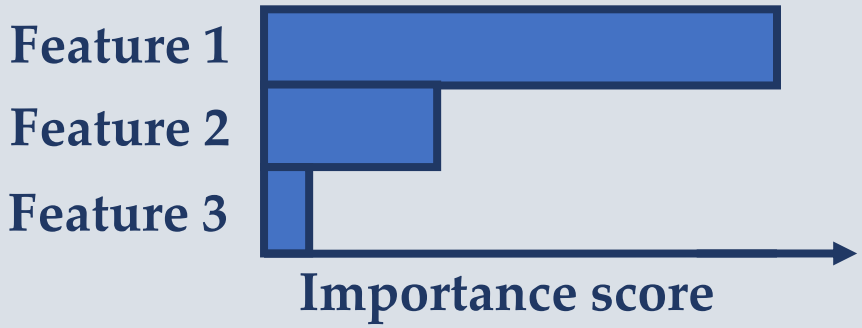
University of Colorado Boulder

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

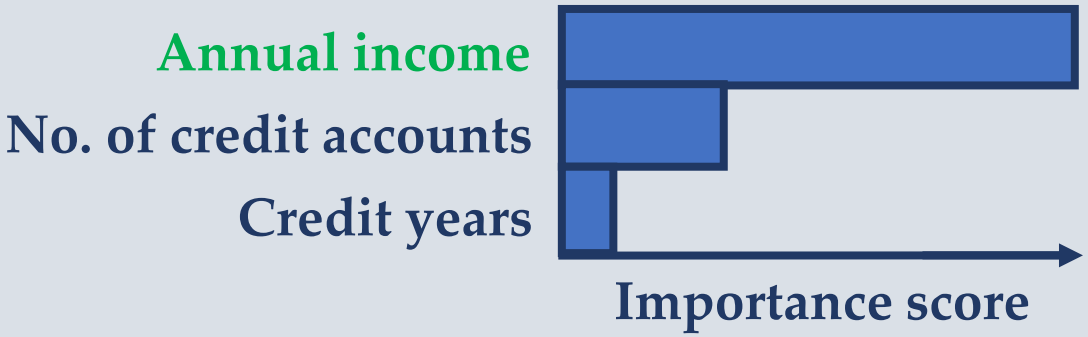


But there is an important problem...

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

Bank → Loan distribution algorithm → Loan granted / Loan denied

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

Feasibility + **Diversity**

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

Wachter et al (2017)

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

Russell (2017)
Mixed integer programming
Works only for linear ML models

General optimization framework

Diverse counterfactual explanations

Loss to get desirable outcome

Loss to ensure proximity to original input

Loss to provide diverse explanations

$$\boldsymbol{C(x)} = \underset{c_1,\ldots,ck}{\arg\min} \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{yloss}(f(c_i), y) + \frac{\lambda_1}{k} \boldsymbol{dist}(c_i, x) - \lambda_2 \boldsymbol{dpp\_diversity}(c_1, .., c_k)$$

k – no. of counterfactuals

λ_1 and λ_2 – loss-balancing hyperparameters

$$dpp\_diversity = det(K),$$

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

Practical considerations

$$\boldsymbol{C(x)} = \arg\min_{c_1,\ldots,ck} \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{yloss}(f(c_i), y) + \frac{\lambda_1}{k}\boldsymbol{dist}(c_i, x) - \lambda_2 \boldsymbol{dpp\_diversity}(c_1, .., c_k)$$

- 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
(Diverse Counterfactual Explanations)
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

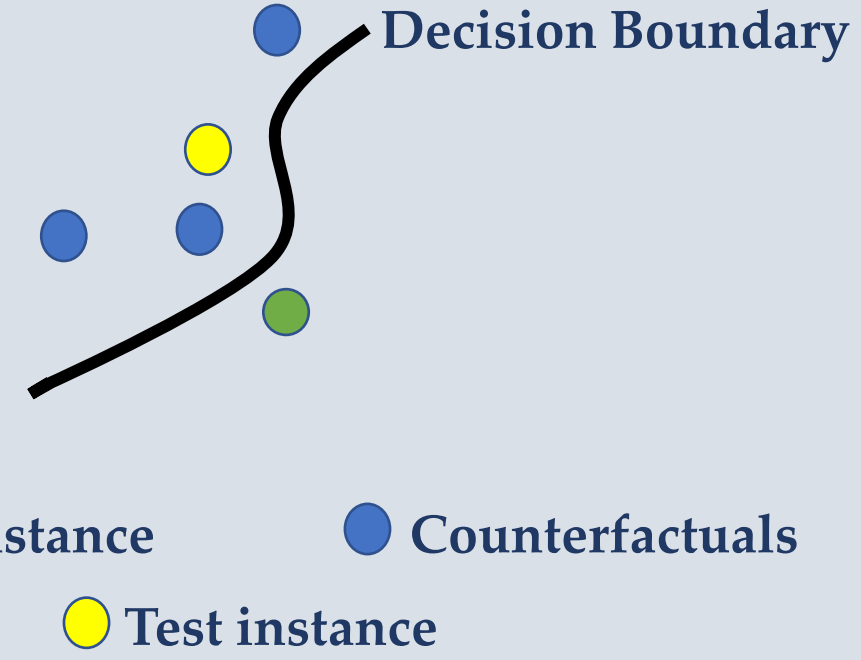
➢ Metrics for comparison

- ❑ Validity : $\frac{|\{unique\ instances\ in\ C\ s.t. f(c)>0.5\}|}{k}$
- ❑ Proximity : $1 - \frac{1}{k}\sum_{i=1}^{k} dist_{cat(c_i,x)}$
- ❑ 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 <u>not the same</u> as CF generation metrics

➢ Approximate local decision boundary

Can CF explanations help users "extrapolate" the local decision boundary of the ML model?



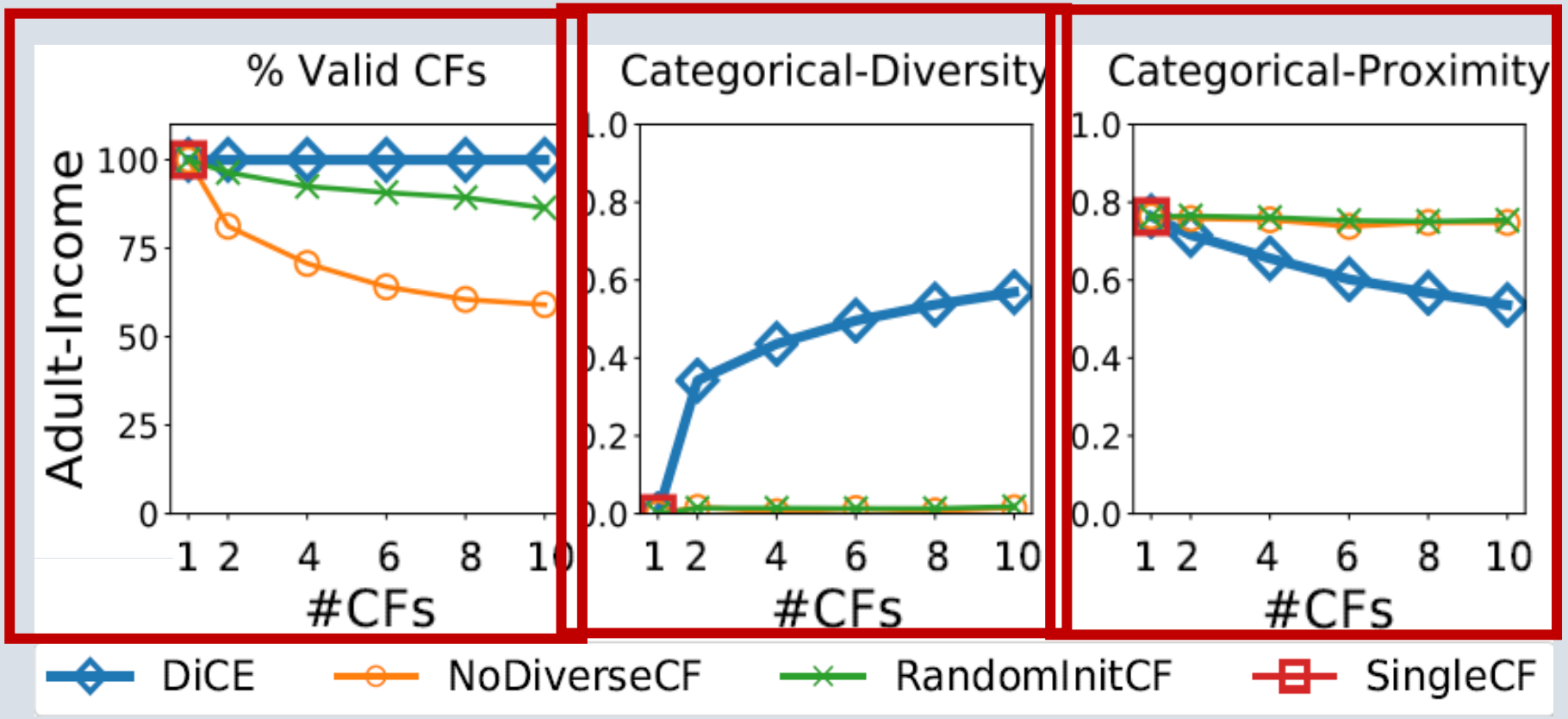
Results: comparing CF-based methods

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

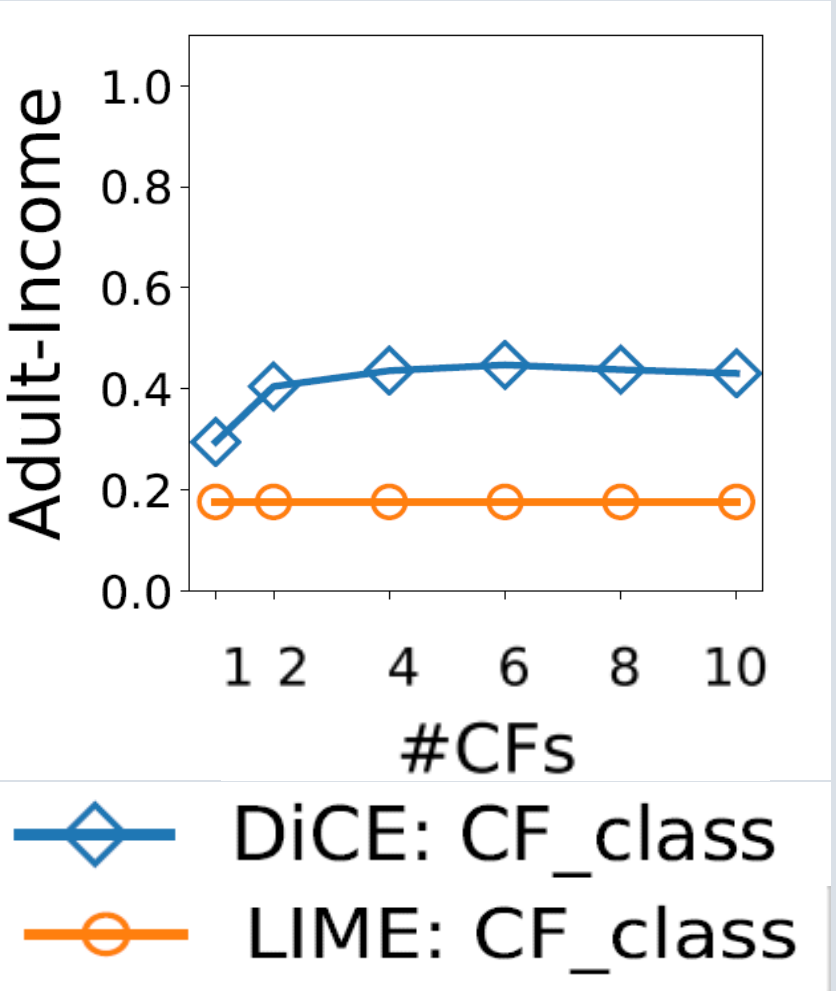


Explaining linear ML models [compared to Russel (2019) and baselines]

- 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**