DiCE FOR ML

Diverse **C**ounterfactual **E**xplanations **for ML**

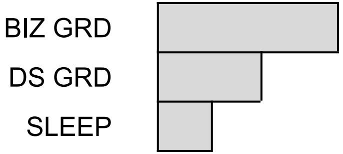
LT3 - Barajas, Fuentebella, Gaspar, Jayme, Ramos, & Tanjangco

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

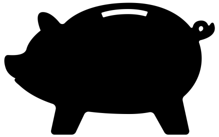
Background: Imagine you are at high risk of failing your Term 3 Grades...

School

Model



Explanation through feature importance



Will not graduate on time

“You may not graduate on time because of your BIZ GRD...”

What should you do next to graduate on time?

Background: Imagine you are at high risk of failing your Term 3 Grades...

School

→

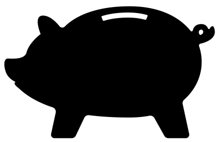
Model

→

Counter-factuals

Explanation through examples

→



Will not graduate on time

"You will graduate on time if you increase your BIZ GRD by 80%..."

Counterfactual examples show how to obtain a different prediction

Counterfactual explanations should satisfy two properties:

feasibility of choices

relative ease of doing the actions (e.g. education, loan, years of work experience)

diversity in choices

it can provide/recommend different actions/scenarios depending on constraint

Counterfactual Engine

Can We Access the Model?

YES	NO
MODEL SPECIFIC	MODEL AGNOSTIC
WHITE BOX APPROACH	BLACK BOX APPROACH

PAPER(SHORTENED NAME)	AUTHORS
CFs Without opening the Blackbox	Watcher et. al.
Contransive Explanation Model	Dhurandar et. al.
Contransive Adversarial Examples	Moore et. al.
Diverse CounterFactual Explanations	**Mothial et. al.**

PAPER(SHORTENED NAME)	AUTHORS
Mod. Agnostic CFs (MACEM)	Dhurandar et. al.
CertifAI	Sharma et. al.
Foil Trees	Van der Waa et. al.

Counterfactual Engine: how are counterfactuals generated?

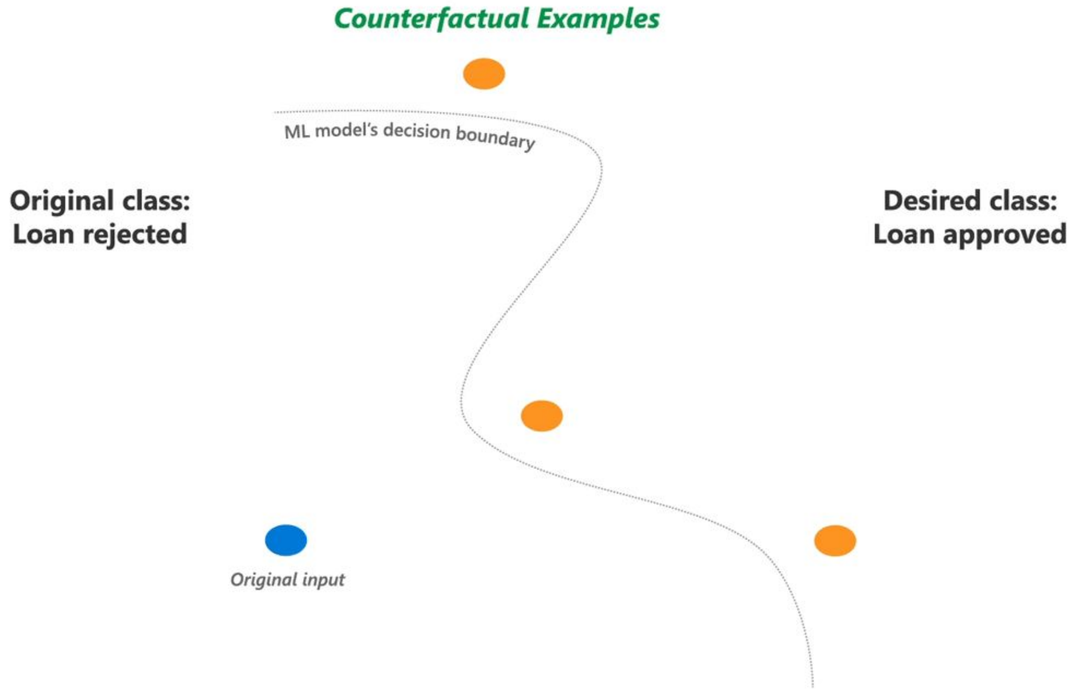
AGE= 71	BMI = 35	GENDER=M	....	STROKE

RANDOM SAMPLING

AGE= 71	BMI = 20	GENDER=M	....	NO STROKE

Model-Agnostic

The Rashomon Effect



Mothilal, R. K., Sharma, A., & Tan, C. (2019). Diverse Counterfactual Explanations (DiCE) for ML. GitHub Repository. Retrieved 15 November 2021 from https://github.com/interpretml/DiCE

Counterfactuals: desirable properties

- **Actionability:**

Users should be able to make the changes indicated by counterfactuals

- **Feasibility**
  - Proximity
  - User constraints
  - Sparsity
  - Causal constraints

+

Diversity

Russell(2017)

Mixed linear programming

Wachter et. al (2017)

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

General Optimization Framework

Diverse counterfactual explanations

Loss to get the desirable outcome

Loss to ensure proximity to the original input

Loss to provide diverse explanation

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

$$dpp\_diversity = \det\left(\frac{1}{1 + dist(c_i, c_j)}\right)$$

k – number of counterfactuals

λ_1 *and* λ_2 – loss – balancing hyperparameter

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

Practical considerations

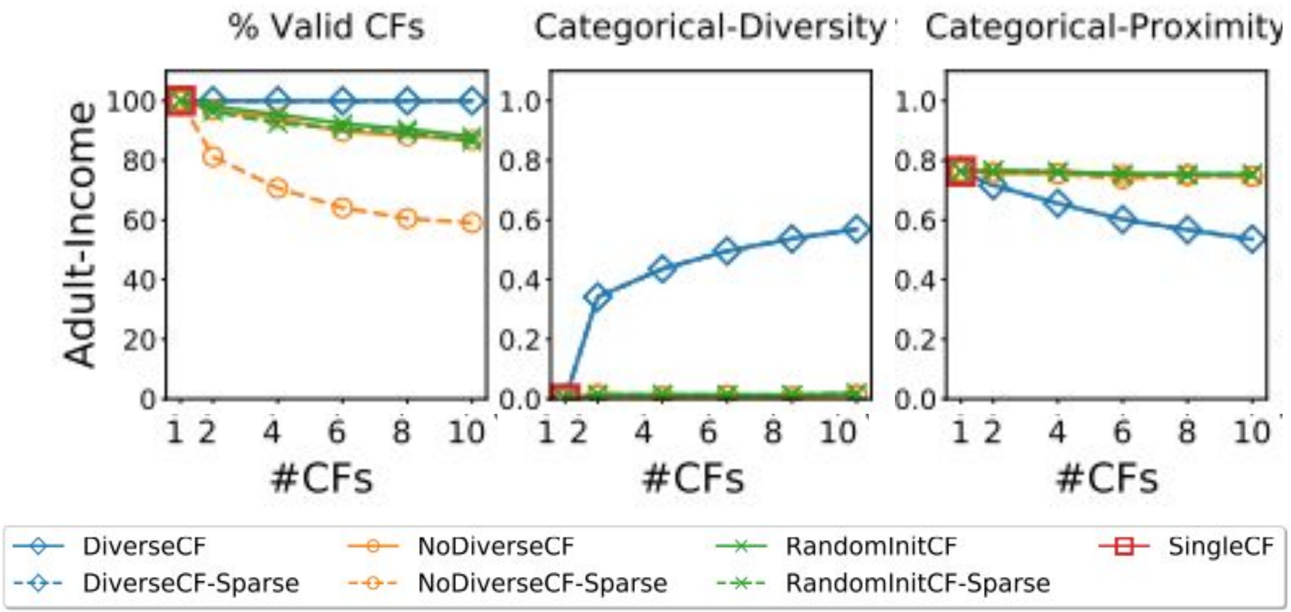
$$C(x) = \arg \min_{c_1,\ldots,c_k} \frac{1}{k} \sum_{i=1}^{k} yloss(f(c_i), y) + \frac{\lambda_1}{k} dist(c_i, x) - \lambda_2 dpp\_diversity(c_1, \ldots, c_k)$$

- Incorporate additional feasibility properties
  - Sparsity
  - User Constraint
- Choice of yloss - **hinge** loss
- Binary classification only

Python Library: DiCE
https://github.com/microsoft/DiCE

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Diverse Counterfactual Explanations (DiCE) for ML. GitHub Repository. Retrieved 15 November 2021 from https://github.com/interpretml/DiCE

Evaluation: Quantitative



Validity - # of valid CFs (*should be high*)

Proximity - ease of adopting a change (*should be high*)

Diversity - suggested changes (*should be high*)

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

Evaluation: Qualitative

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	—

Counterfactuals can be evaluated one-by-one

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

Demonstration: UPCAT College Exam Prediction

```
#prepare the data
dataset = X_new
target = dataset["Target"]

train_dataset, test_dataset, y_train, y_test = train_test_split(dataset,
                                                    target, test_size=0.25,
                                                    random_state=0)
x_train = train_dataset.drop('Target', axis=1)
x_test = test_dataset.drop('Target', axis=1)

# setup the data
d = dice_ml.Data(dataframe=train_dataset,
                 continuous_features=['Eng7', 'Eng8','Eng9','Math7',
                                      'Math8','Math9', 'Sci7','Sci8', 'Sci9',
                                      'GWA7', 'GWA8','GWA9', 'UP','SA', 'C1',
                                      'C2'], outcome_name='Target')

#define the algorithm
clf = Pipeline(steps=[('classifier', RandomForestClassifier())])
model = clf.fit(x_train, y_train)

#run dice
m = dice_ml.Model(model=model, backend="sklearn")
# Using method=random for generating CFs
exp = dice_ml.Dice(d, m, method="random")

#generate and visualize the counterfactuals
e1 = exp.generate_counterfactuals(x_test[0:10], total_CFs=3,
                                  desired_class="opposite")
e1.visualize_as_dataframe(show_only_changes=True)
```

We ran the code using the UPCAT College Exam Prediction dataset...

Demonstration: UPCAT College Exam Prediction Example

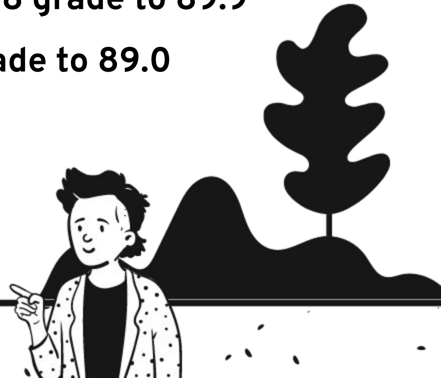
	Eng7	Eng8	Eng9	Math7	Math8	Math9	Sci7	Sci8	Sci9	UP	IQ	Target
0	77	73.0	79	85	80	71.0	73	88	78	1	55.0	0

```
Diverse Counterfactual set (new outcome: 1.0)
```

	Eng7	Eng8	Eng9	Math7	Math8	Math9	Sci7	Sci8	Sci9	UP	IQ	Target
0	-	-	-	-	-	76.3	-	-	-	-	-	1.0
1	-	-	-	-	-	-	59.0	-	-	-	-	1.0
2	-	89.9	89.0	-	-	-	-	-	-	-	-	1.0

For this individual to pass, they can do any of the following:

- Increase Math9 grade to 76.3
- Lower Sci7 grade to 59.0
- Increase Eng8 grade to 89.9 and Eng9 grade to 89.0

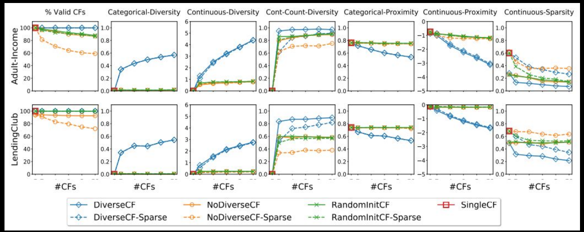


Summary: Contribution to the field

Counterfactual explanations should satisfy two properties:

feasibility of choices | **diversity in choices**

and was demonstrated and evaluated:



quantitatively

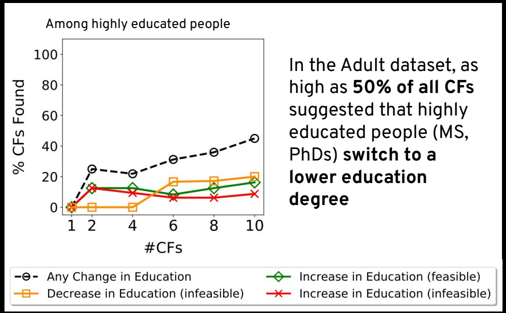
Adult	HrsWk	Education	Occupation
Original input (outcome: <=50K)	45.0	HS-grad	Service
Counterfactuals (outcome: >50K)	—	Masters	—
	—	Doctorate	—
	33.0	—	White-Collar
	57.0	Prof-school	—

qualitatively

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

Summary: Further work

Limitation



May lead to causal infeasibility

Areas for extension

Compare with SHAP and other interpretable methods

"Do DiCE-produced counterfactuals provide better explanations than past ML interpretability approaches?"

Behavioral study

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

DiCE FOR ML

Diverse **C**ounterfactual **E**xplanations **for ML**

LT3 - Barajas, Fuentebella, Gaspar, Jayme, Ramos, & Tanjangco

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850