Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Towards a Synergistic human-machine Interaction and Collaboration

Roberto Pellungrini

Scuola Normale Superiore

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Three Learning Paradigms

Human Oversight

Learning to Abstain

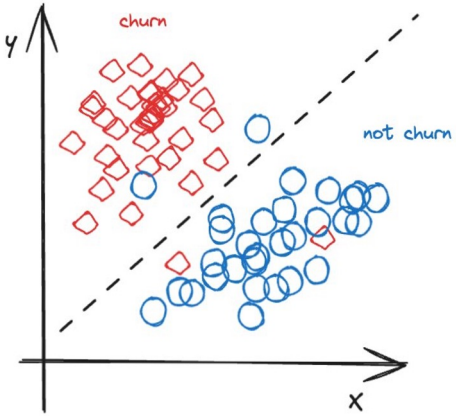
Learning Together

Punzi C., Pellungrini R., Setzu M., Giannotti F., Pedreschi D.,. **AI, Meet Human: Learning Paradigms for Hybrid Decision Making Systems**, Under revision at ACM Computing Surveys, 2023.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Ensemble Counterfactual Explanations for Churn Analysis

Customer churn is the percentage of customers who stopped purchasing a company business's products or services during a certain period of time



Evaluation from questions

***Q.1**: How minimal are the changes required to retain potentially churning customers?*

$$\text{Proximity}_{L2} = \sqrt{\frac{m-h}{m}\sum_{i\in\text{cont}}(x'_i - x_i)^2 + \frac{h}{m}\sum_{j\in\text{cat}}\delta(x'_j, x_j)}$$

***Q.2**: Is the counterfactual explanation similar to a non churning customer in the data and thus justifiable to the customer?*

$$\text{Plausibility} = \sqrt{\sum_{i=1}^{d}(x'_i - x_{\text{nearest},i})^2}$$

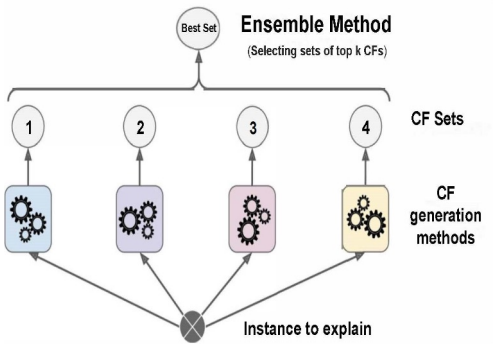
***Q.3**: Does the counterfactual explanation suggests changes that are easy for the churn officer to propose?*

$$\text{Sparsity} = \frac{\sum_{i=1}^{n}(x'_i \neq x_i)}{n}$$

***Q.4**: Do the counterfactuals produced provide different courses of action for the churn officer?*

$$\text{Diversity} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq k}d(x_j, x_k)$$

Putting things together



$$\text{Evaluation Score} = \begin{aligned} &+ w_1 \times \text{Proximity}_{\text{avg;N}} \\ &+ w_2 \times \text{Sparsity} \\ &+ w_3 \times \text{Plausibility}_{\text{N}} \\ &+ w_4 \times (1 - \text{Diversity}_{\text{N}}) \end{aligned}$$

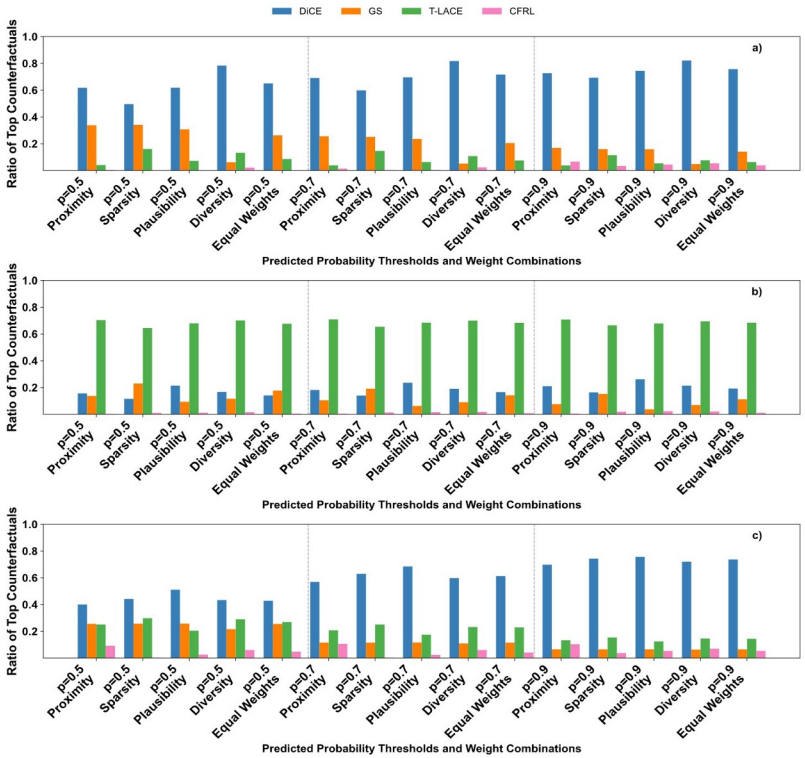
- Four Counterfactual Generation Methods
- Ensemble of Counterfactual Sets
- Linear combination Evaluation Score
- Selection of top K Counterfactual Examples

- *Allows a diverse sets of counterfactual explanations for a given instance*
- *Weights can be tweaked by the user to select the counter-exemplars with the desired properties for user segmentation*

Guidotti, R., Ruggieri, S. (2021). **Ensemble of Counterfactual Explainers**. In: Soares, C., Torgo, L. (eds) Discovery Science. DS 2021. Lecture Notes in Computer Science(), vol 12986. Springer, Cham. https://doi.org/10.1007/978-3-030-88942-5_28

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Results



Original Instance

CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Churn
650	1	30	6	0	1	0	0	67997	1

Counterfactual Instances

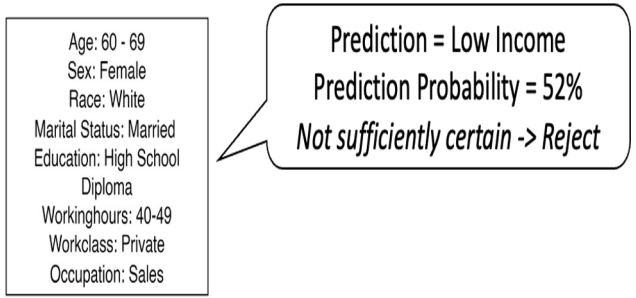
CF_1: Tenure $+1 \rightarrow$ Churn $= 0$
CF_2: CreditScore -157, NumOfProducts $+1$, EstimatedSalary -2181 $\rightarrow$ Churn $= 0$
CF_3: HasCrCard $+1 \rightarrow$ Churn $= 0$
CF_4: CreditScore -300, NumOfProducts $+1$, EstimatedSalary -2807 $\rightarrow$ Churn $= 0$
CF_5: Tenure $+3$, EstimatedSalary -67997 $\rightarrow$ Churn $= 0$

Tonati S., Pellungrini, R., Di Vece, M., Giannotti, F., (2024). **Ensemble Counterfactual Explanation for Churn Analysis.** Accepted at Discovery Science 20204

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Interpretable & Fair Mechanisms for Abstaining Classifiers (IFAC)

- Classifiers rejecting instances based on uncertainty of predictions (depending on coverage)
- Common approach: prediction probability as proxy for certainty



Can we extend the selective classification framework?

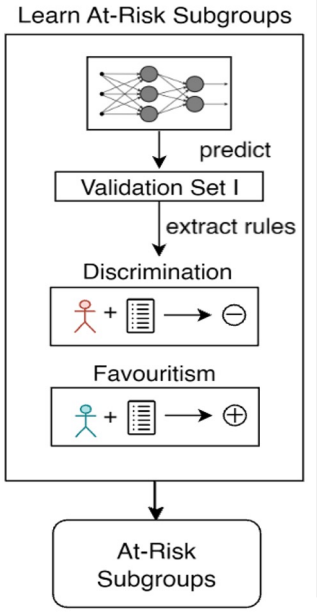
Classifier rejecting instances based on unfairness of predictions

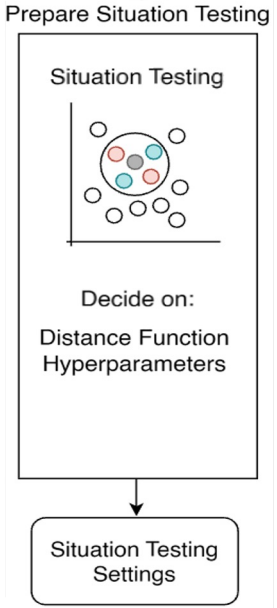
Possibilities:

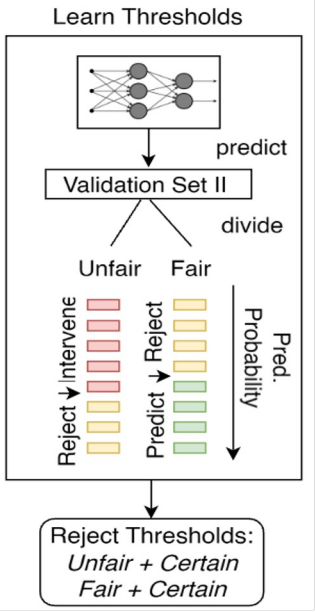
- Pass rejected instances on to human-in-the loop to examine (EU AI Act)
- Give explanation behind rejection, for better informed fairness judgements

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

How IFAC is trained







1. Train Black Box Classifier on Data (BB)
2. Learn Discriminatory Associations in BB
3. Decide Parameters for Local Fairness Check (Situation Testing algorithm)
4. Estimate (un)certainty thresholds

	Certain	Uncertain
Fair	Predict	Reject
Unfair	Reject	Intervene

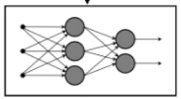
t_fair_certain, t_unfair_certain

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

How IFAC works

Base Classifier Prediction

Age: 60 - 69
Sex: Female
Race: White
Marital Status: Married
Education: High School Diploma
Workinghours: 40-49
Workclass: Private
Occupation: Sales



Predicted Label: Low Income
Prediction Probability: 0.7417

Global Fairness Check

At-Risk of Discrimination

sex = Female AND age = 60 - 69 AND occupation = Sales

race = Black AND education = Master AND age = 50-59

sex = Female AND race = Other AND occupation = Engineering

At-Risk of Favouritism

sex = Male AND race = White AND education = Bachelor AND workinghours = More than 50

sex = Male AND race = White AND education = Master

Local Fairness Check

Predictions for similar instances 'White & Male'

M. Status	Education	W.Hours	W.Class	Pred.
Married	High School	40-49	Private	High
Married	High School	40-49	Private	High
Divorced	High School	40-49	Private	Low

Predictions for similar instances NOT 'White & Male'

M. Status	Education	W.Hours	W.Class	Pred.
Married	High School	40-49	Private	Low
Widow	High School	40-49	Private	Low
Married	High School	30-39	Private	Low

Individual Discrimination Score = (2/3) - (0/3) = 2/3

Global + Local Check Fail

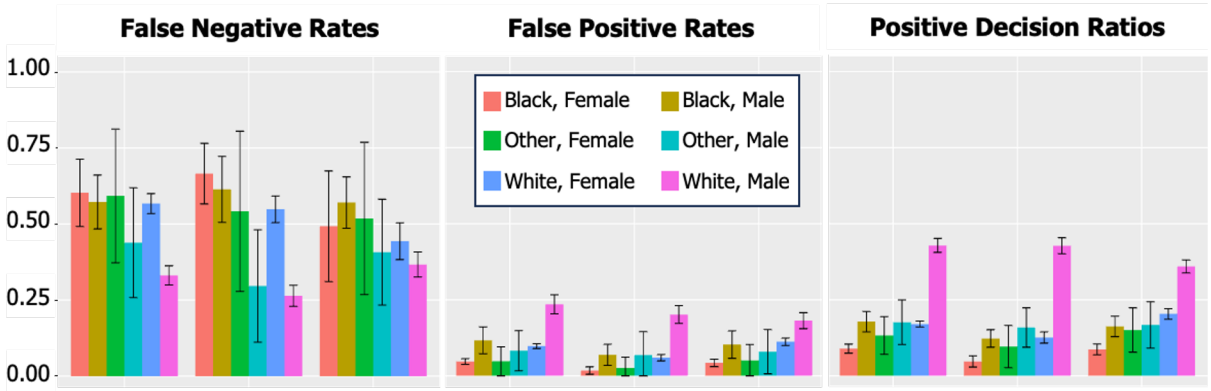
Certainty Check

Prediction Probability = 0.7417

Probability lays **above** certainty threshold for unfair predictions

Unfair + Certain Prediction: REJECT!

Fairness improvements



	FC	USC	FSC	FC	USC	FSC	FC	USC	FSC
Range	0.27	0.40	**0.20**	0.19	0.18	**0.14**	0.34	0.38	**0.27**
Std.	0.11	0.17	**0.08**	0.07	0.07	**0.05**	0.12	0.13	**0.09**

Explanation

(sex = Female AND race = Black AND age = 30-39 AND occupation = Management) -> low income

Confidence: 1.000, SLift: 0.658

age: 30-39 **marital status:** Married **education:** *High School* **workinghours:** *20-39* **workclass**: private **occupation:** Management **race**: Black **sex**: Female	**High Income Rates** **Similar white men: 6/10** **Similar non white men: 2/10** - 4/6 of white men with high income have a Bachelor - 6/6 work at least 40-49 hours
age: 30-39 **marital status:** Married **education**: Bachelor Degree **workinghours**: 40-49 **workclass**: private **occupation**: Management **race**: Black **sex:** Female	**Similar white men: 9/10** **Similar non white men: 3/10** - all of them share instance's education, workinghours, marital status workclass

- Two instances falling under same possibly discriminated subgroup
- Human-in-the loop gets to make *final* decision

Lenders, D., Pugnana, A., Pellungrini, R., Calders, T., Pedreschi, D., Giannotti, F. (2024). **Interpretable and Fair Mechanisms for Abstaining Classifiers**. In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds) Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2024. Lecture Notes in Computer Science(), vol 14947. Springer, Cham. https://doi.org/10.1007/978-3-031-70368-3_25

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Future Works

Interactive explanations

Explainable mechanism for abstention

Datasets for human deferral