Optimal Counterfactual Explanations in Tree Ensembles

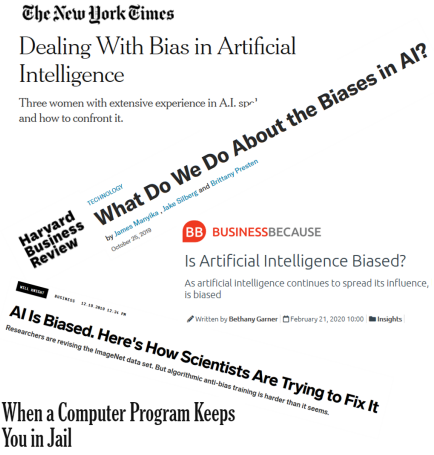
Axel Parmentier[1], Thibaut Vidal[2,3]

[1] Cermics, École des Ponts Paristech
[2] CIRRELT & SCALE-AI Chair in Data-Driven Supply Chains, MAGI, Polytechnique Montreal, Canada
[3] Department of Computer Science, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil

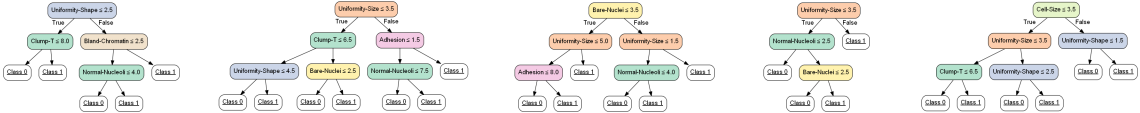
Image Credit: Hitchhikers Guide to the Galaxy

Sensitive applications of ML require Transparency and Explainability

- **Machine learning** applied to high stakes decisions:
  - Recurrence predictions in medicine
  - Credit default risk evaluations
  - Even in contexts where it should not be applied in current form (e.g., bail decisions in criminal justice...)
- **Critical decisions** $\Rightarrow$ Right to have explanations and recourse, i.e., "what can I do to change the outcome".
- **Counterfactual explanations:** contrastive arguments of the type: "To obtain this loan, you need \$40,000 of annual revenue instead of the current \$30,000".
  - Ideally, good counterfactual explanations provide the "smallest" set of changes of the features (or actions) needed to achieve the desired class,
  - Bound by additional constraints imposing plausibility and actionability.



Explanations in Tree Ensembles



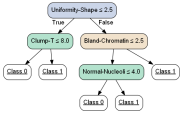
Counterfactual Search

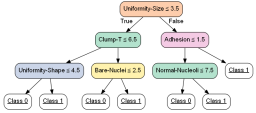
Given an origin point $\hat{\mathbf{x}}$ and a desired prediction class c^*, searching for a plausible and actionable counterfactual explanation consists in locating a new data point $\mathbf{x} \in \mathcal{X}$ that solves the following problem:

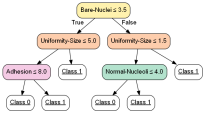
$$\begin{aligned} \min \quad & f_{\hat{\mathbf{x}}}(\mathbf{x}) \\ \text{s.t.} \quad & F_{\mathcal{T}}(\mathbf{x}) = c^* \\ & \mathbf{x} \in X^{\mathrm{P}} \cap X^{\mathrm{A}} \end{aligned}$$

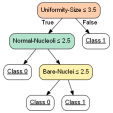
- Finding counterfactual explanations in **tree ensembles** is notably difficult:
  - Function $F_{\mathcal{T}}(\mathbf{x})$ has a number of pieces that grows as the *product* of the number of leaves of the trees [7].
  - Changing any feature impacts the trajectory in all trees $\Rightarrow$ searching for the right combination of leaves
  - Non-convex, non differentiable decision function, NP-hard optimization problem.

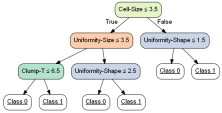
HEURISTIC vs OPTIMAL Explanations











- Current **HEURISTIC** explanation algorithms, e.g., Feature Tweaking (FT – [5]) regularly produce suboptimal solutions
  - Largely overshooting the actions (up to 31.7× in our experiments) needed to achieve the desired outcome.
  - Unstable solution quality, widely varying between different subjects and subject groups...
  - ***Is this transparent and fair?***

- **OPTIMAL** counterfactual search through mixed-integer linear programming (MILP) provides explanations grounded on a mathematical definition, independently of the search algorithm
  - The flexibility of the modeling framework permit to seamlessly include a wide diversity of metrics, objectives and constraints
  - As seen in this study, is possible to achieve optimal results within seconds

OCEAN – Optimal Counterfactual Explanations

- MILP = solution of a problem represented as a set of linear equations, in which some variables are restricted to the integer domain, under an objective measuring how difficult it is to act on the different features.
- Solved to optimality with a branch-and-cut solver (Gurobi)
- Our model has several desirable characteristics that permit an efficient solution:
  - **logarithmic number of integer variables** (not linear as in [1, 2]);
  - **tighter linear relaxation** than previous models $\Rightarrow$ improves branch-and-cut performance
- The approach is very flexible:
  - applicable to **heterogeneous data** with numerical, ordinal, categorical and binary features;
  - large **variety of objectives**: l_0, l_1 and l_2 norm and extensions thereof;
  - additional **actionability** and **plausibility** constraints.

Sample Flows

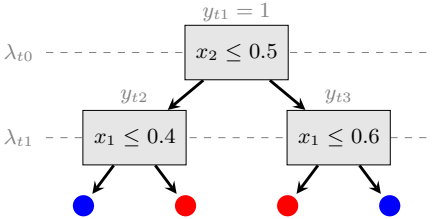
The $\boldsymbol{\lambda}$ variables represent the branch decisions for each tree t at each layer d. The $\mathbf{y}$ variables represent the flows of the counterfactual example through each tree.

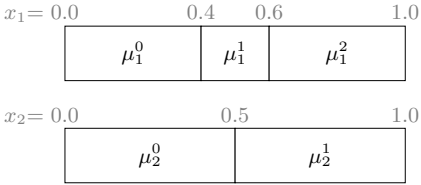
$$
\begin{aligned}
& y_{t1} = 1 && t \in \mathcal{T} \\
& y_{tv} = y_{tl(v)} + y_{tr(v)} && t \in \mathcal{T}, v \in \mathcal{V}_t^I \\
& \sum_{v \in \mathcal{V}_{td}^I} y_{tl(v)} \leq \lambda_{td} && t \in \mathcal{T}, d \in \mathcal{D}_t \\
& y_{tv} \in [0,1] && t \in \mathcal{T}, v \in \mathcal{V}_t^I \cup \mathcal{V}_t^I \\
& \lambda_{td} \in \{0,1\} && t \in \mathcal{T}, d \in \mathcal{D}_t.
\end{aligned}
$$

Numerical Features

The $\boldsymbol{\mu}$ variables represent the levels of numerical features as ordered simplices, also connecting them with the variables representing the branch choices.

$$
\begin{aligned}
& \mu_i^{j-1} \geq \mu_i^j && j \in \{1, \ldots, k_i\} \\
& \mu_i^j \leq 1 - y_{tl(v)} && j \in \{1, \ldots, k_i\}, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
& \mu_i^{j-1} \geq y_{tr(v)} && j \in \{1, \ldots, k_i\}, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
& \mu_i^j \geq \epsilon y_{tr(v)} && j \in \{1, \ldots, k_i\}, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
& \mu_i^j \in [0,1] && j \in \{0, \ldots, k_i\}
\end{aligned}
$$





OCEAN – Optimal Counterfactual Explanations

Categorical Features

The $\boldsymbol{\nu}$ variables represent the possible categories:

$$\begin{aligned}
&\nu_i^j \leq 1 - y_{tl(v)} && j \in C_i, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
&\nu_i^j \geq y_{tr(v)} && j \in C_i, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
&\nu_i^j \in \{0,1\} && j \in C_i \\
&\sum_{j \in C_i} \nu_i^j = 1
\end{aligned}$$

Ordinal Features

The $\boldsymbol{\omega}$ variables represent the relevant levels for the ordinal features (only those appearing in some splitting hyperplanes):

$$\begin{aligned}
&\omega_i^{j-1} \geq \omega_i^j && j \in \{2, \ldots, k_i - 1\} \\
&\omega_i^j \leq 1 - y_{tl(v)} && j \in \{1, \ldots, k_i - 1\}, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
&\omega_i^j \geq y_{tr(v)} && j \in \{1, \ldots, k_i - 1\}, t \in \mathcal{T}, v \in \mathcal{V}_{tij}^I \\
&\omega_i^j \in \{0,1\} && j \in \{1, \ldots, k_i - 1\}
\end{aligned}$$

Binary Features

Simply done through binary $\mathbf{x}$ variables:

$$\begin{aligned}
&x_i \leq 1 - y_{tl(v)} && t \in \mathcal{T}, v \in \mathcal{V}_{ti}^I \\
&x_i \geq y_{tr(v)} && t \in \mathcal{T}, v \in \mathcal{V}_{ti}^I \\
&x_i \in \{0,1\}
\end{aligned}$$

The domain of the variables representing the features can be relaxed to the continuous interval $[0, 1]$ while retaining integrality of the linear-relaxation solutions (with the simplex algorithm).

OCEAN – Optimal Counterfactual Explanations

For **binary**, **categorical**, or **ordinal** features, we can freely set a weight for each discrete choice in the objective. For **continuous** numerical features, we can proceed as follows:

Objective for numerical features

$$l_0 : \begin{cases} f_0^{\mathrm{N}}(\boldsymbol{\mu}) = \sum_{i \in I_{\mathrm{N}}} (c_i^- z_i^- + c_i^+ z_i^+) & \\ z_i^- \geq 1 - \mu_i^{j-1}, z_i^+ \geq \mu_i^j & i \in I_{\mathrm{N}}, j = \hat{j}_i \\ z_i^- \in \{0,1\}, z_i^+ \in \{0,1\} & i \in I_{\mathrm{N}} \end{cases}$$

$$l_1 : \begin{cases} f_1^{\mathrm{N}}(\boldsymbol{\mu}) = \sum_{j=0}^{k_i} (\phi_i^{j+1} - \phi_i^j)\mu_i^j & \\ \text{with parameter } \phi_i^j = c_i^- \max(\hat{x}_i - x_i^j, 0) & \\ \qquad + c_i^+ \max(x_i^j - \hat{x}_i, 0) & \end{cases}$$

Achieving the desired counterfactual class c^* though majority vote can be expressed as:

$$z_c = \sum_{t \in \mathcal{T}} \sum_{v \in \mathcal{V}_t^L} w_t p_{tvc} y_{tv} \qquad c \in \mathcal{C}$$

$$z_{c^*} > z_c \qquad c \in \mathcal{C}, c \neq c^*$$

OCEAN – Optimal Counterfactual Explanations

Many additional constraints related to actionability and plausibility can be seamlessly integrated into the model:

Domain Knowledge	**Constraints**
Fixed features	$x_i = \hat{x}_i,\ \mu_i = \hat{\mu}_i,\ \nu_i = \hat{\nu}_i$
Monotonic features	$x_i \geq \hat{x}_i,\ \mu_i \geq \hat{\mu}_i,\ \nu_i \geq \hat{\nu}_i$
Known **linear relations** between features (i.e., joint actionability – Venkatasubramanian and Alfano 6)	$A(x_i - \hat{x}_i) \leq \mathbf{b}$
Known **logical implications** between features, Example for binary features $(x_1 = \text{True}) \Rightarrow (x_2 = \text{True})$ Example for categorical features $x_1 \in \{\text{Cat1}, \text{Cat2}\} \Rightarrow x_2 \in \{\text{Cat3}, \text{Cat4}\}$	 $x_2 \geq x_1$ $\nu_2^3 + \nu_2^4 \geq \nu_1^1 + \nu_1^2$
Resource constraints (e.g., time) as modeled by additional functions $g_i(\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\mu})$	$g_i(\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\mu}) \leq b_i$

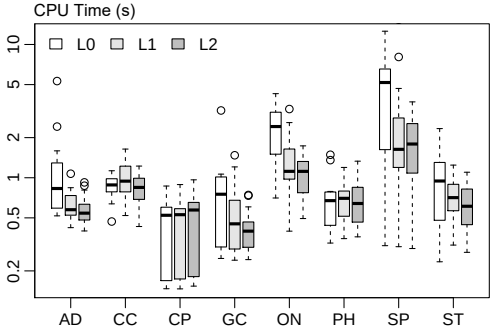
Experimental Setup

- We used heterogeneous datasets coming from a wide range of applications, with up to 45222 samples and 57 features.
  - Divided each data set into 80% training and 20% test
  - For each data set, trained a random forest (RF) with 100 trees and maximum depth of 5, and selected 20 origin samples with negative outcome for the counterfactual explanations
  - Saved/Serialized all the RF and samples for a fair comparison between different counterfactual explanation methods

Data set	n	p	p_N	p_B	p_C	Src.
AD: Adult	45222	11	5	2	4	UCI
CC: Credit Card Default	29623	14	11	3	0	UCI
CP: COMPAS	5278	5	2	3	0	ProPublica
GC: German Credit	1000	9	5	1	3	UCI
ON: Online News	39644	47	43	2	2	UCI
PH: Data Phishing	11055	30	8	22	0	UCI
SP: Spambase	4601	57	57	0	0	UCI
ST: Students Performance	395	30	13	13	4	UCI

Computational Experiments – Performance and Optimality

Measuring the time needed to find optimal counterfactual explanations with OCEAN for different objectives and data sets



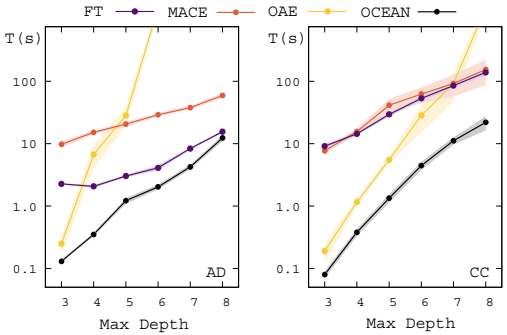
Comparing CPU time and solution quality with other approaches for RF explanations: FT [5], MACE [3], OAE [1, 2]. We use the l_1 objective (which is common to all methods) and the same serialized RFs.

Data	FT		MACE		OAE		OCEAN	
	T(s)	R	T(s)	R	T(s)	R	T(s)	R
AD	3.03	15.9	20.60	1.1	28.37	1.0	1.22	1.0
CC	29.44	10.2	41.25	1.2	5.52	1.0	1.34	1.0
CP	22.68	4.5	15.82	1.0	0.38	1.0	0.52	1.0
GC	16.26	4.8	19.03	1.0	5.08	1.0	1.16	1.0
ON	10.05	31.7	>900	—	>900	—	2.97	1.0
PH	10.95	1.4	>900	—	0.94	1.0	0.52	1.0
SP	NA	—	>900	—	>900	—	2.73	1.0
ST	NA	—	>900	—	69.64	1.0	1.10	1.0

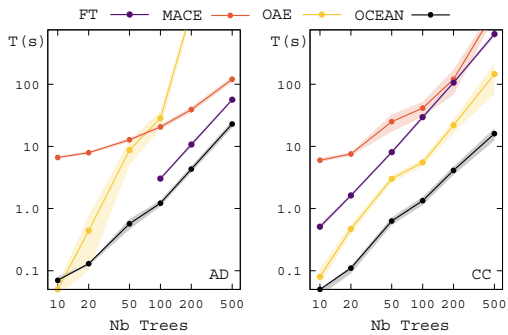
R = Ratio between l_1 distance found & optimum
NA = No counterfactual explanation found
> 900 = Time limit exceeded

Computational Experiments – Scalability

Comparative analysis of CPU time as a function of the maximum depth of the trees. Number of trees fixed to 100:

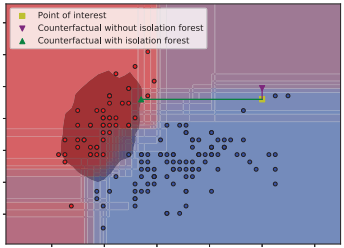


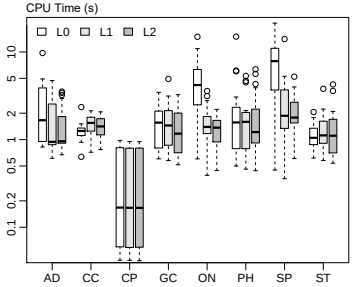
Comparative analysis of CPU time as a function of the number of trees in the ensemble. Maximum depth fixed to 5:



Computational Experiments – Isolation Forests for Plausibility

- Isolation forests [4] are trained to return an outlier score for any sample, inversely proportional to its average path depth within a set of randomized trees.
  - Constraining this average depth to be greater than a threshold δ controls the plausibility of the counterfactual explanation.
  - This is done within the same MILP formulation, with an additional set of constraints representing the IF. We select δ to capture 10% of the training data as an outlier $\Rightarrow$ counterfactual explanation typical of the 90% most common samples for the target class.
  - Computational time remains tractable even with the addition of the IF





Conclusions & Perspectives

- Optimal counterfactual explanations are achievable for most tabular datasets of practical interest
- The flexibility of an appaoch based on MILP gives much-needed flexibility to integrate additional objectives, penalty terms, and constraints related to actionability and plausibility
- Models are still evolving, and likely to need customization for each application at hand
- Further developments could focus on improving performance, but without losing sight of formulation ease and extendability

Thanks !

Contact: thibaut.vidal@polymtl.ca
Data & Open-source code: https://github.com/vidalt/OCEAN
Regular updates: https://twitter.com/vidalthi

References

[1] Cui, Z., W. Chen, W. He, Y. Chen. 2015. Optimal action extraction for random forests and boosted trees. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 179–188.

[2] Kanamori, K., T. Takagi, K. Kobayashi, H. Arimura. 2020. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* 2855–2862.

[3] Karimi, A.-H., G. Barthe, B. Balle, I. Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. Silvia Chiappa, Roberto Calandra, eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 108. PMLR, 895–905.

[4] Liu, F.T., K.M. Ting, Z.-H. Zhou. 2008. Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*. 413–422.

[5] Tolomei, Gabriele, Fabrizio Silvestri, Andrew Haines, Mounia Lalmas. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, 465–474.

[6] Venkatasubramanian, Suresh, Mark Alfano. 2020. The philosophical basis of algorithmic recourse. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 284–293.

[7] Vidal, T., M. Schiffer. 2020. Born-again tree ensembles. Hal Daumé III, Aarti Singh, eds., *Proceedings of the 37th International Conference on Machine Learning*, vol. 119. PMLR, Virtual, 9743–9753.