Robust Counterfactual Explanations for Tree-Based Ensembles

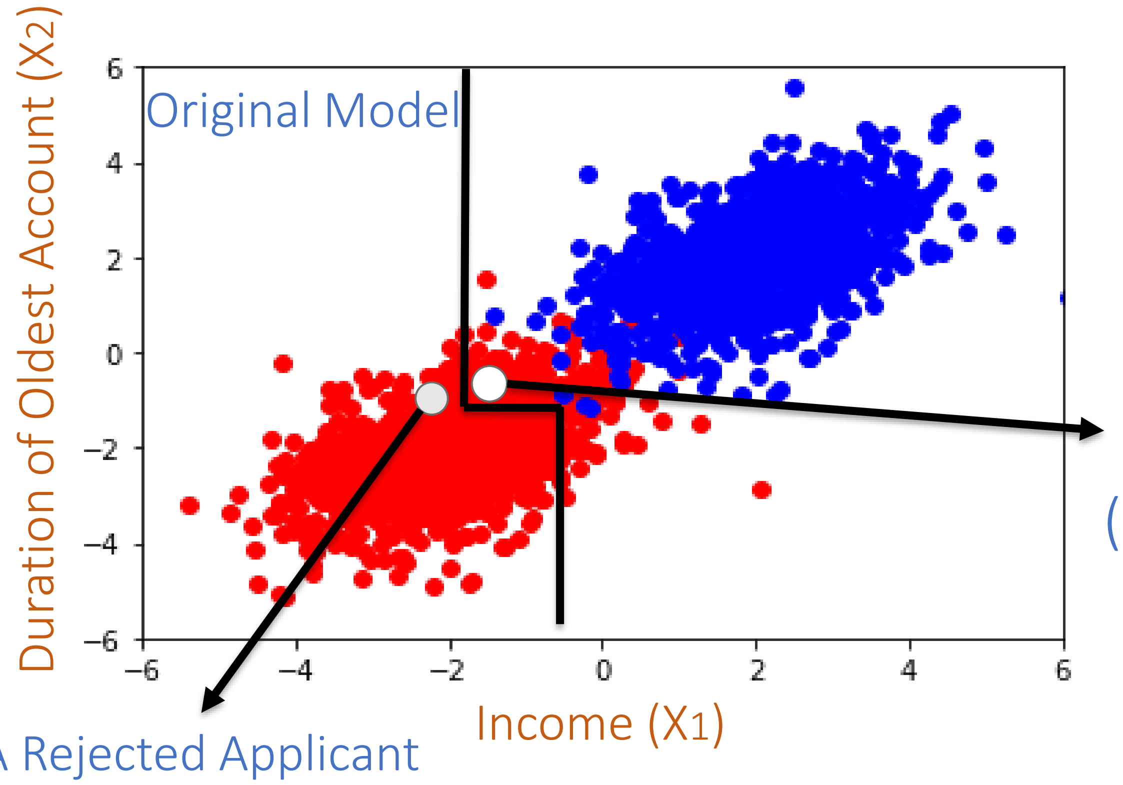
Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, Daniele Magazzeni

JP Morgan AI Research

ICML 2022

Counterfactual Explanations in High-Stakes Applications

Motivation: Reliably guide an applicant on how they can change the model outcome

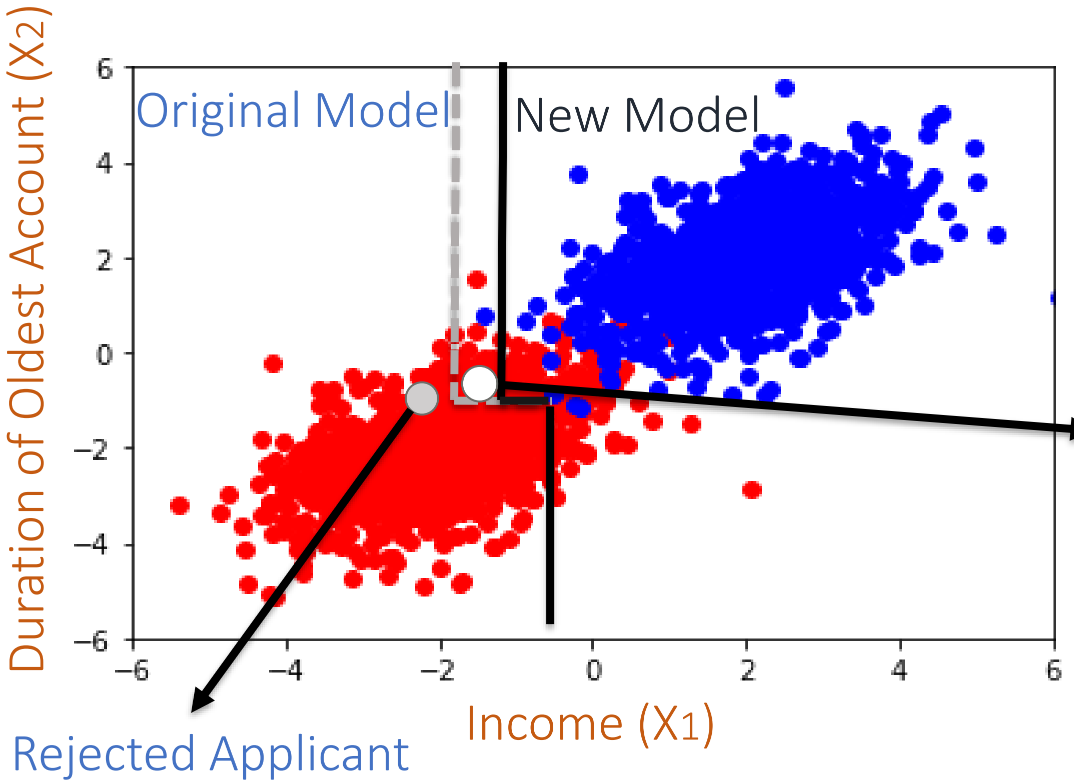


Counterfactual Explanation
(e.g., increase income by 10K)

💡 Find the "closest" point on the other side of the decision boundary

Counterfactual Explanations in High-Stakes Applications

Motivation: Reliably guide an applicant on how they can change the model outcome



Find the "closest" point on the other side of the decision boundary

Counterfactual Explanation (e.g., increase income by 10K)

How do we provide counterfactual explanations that are not only "closest" but also **robust** to model changes?

Problem Statement

Given a data point $x \in \mathcal{X}$ such that $M(x) \leq 0.5$, our **goal** is to find a counterfactual x' with $M(x') > 0.5$ that meets our requirements:

- **Close**, i.e., $||x - x'||_p$ is low
- **Valid** after changes to the model, i.e., $M_{new}(x') > 0.5$
- **Realistic** with respect to the data manifold, i.e., has a better LOF

Problem Statement

Given a data point $x \in \mathcal{X}$ such that $M(x) \leq 0.5$, our **goal** is to find a counterfactual x' with $M(x') > 0.5$ that meets our requirements:

- **Close**, i.e., $||x - x'||_p$ is low
- **Valid** after changes to the model, i.e., $M_{new}(x') > 0.5$
- **Realistic** with respect to the data manifold, i.e., has a better LOF

Related Works:
[Upadhyay et al.'21][Rawal et al.'21][Black et al.'21]

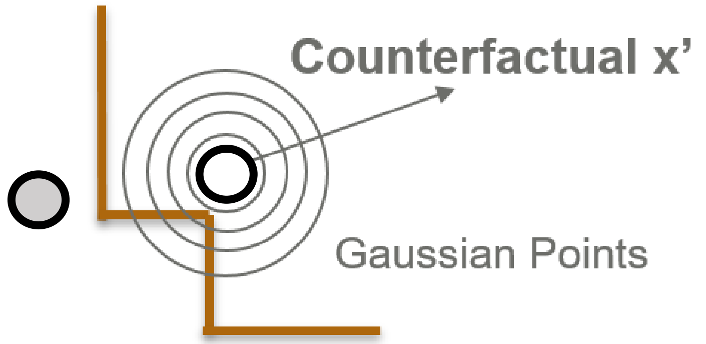
Our focus:
Tree-based models

Contribution 1: Counterfactual Stability

A Novel Measure to Quantify Robustness for Tree-Based Ensembles

$$R_{K,\sigma^2}(x,M) = \frac{1}{K}\sum_{x' \in N_x} M(x') - \sqrt{\frac{1}{K}\sum_{x' \in N_x}\left(M(x') - \frac{1}{K}\sum_{x' \in N_x} M(x')\right)^2}$$

where N_x is a set of K points from the distribution $N(x,\sigma^2)$

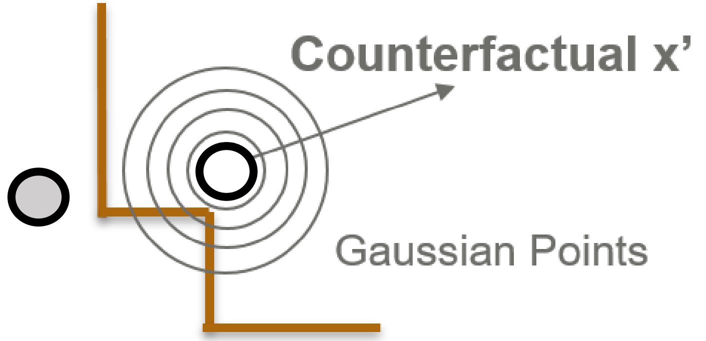


Contribution 1: Counterfactual Stability

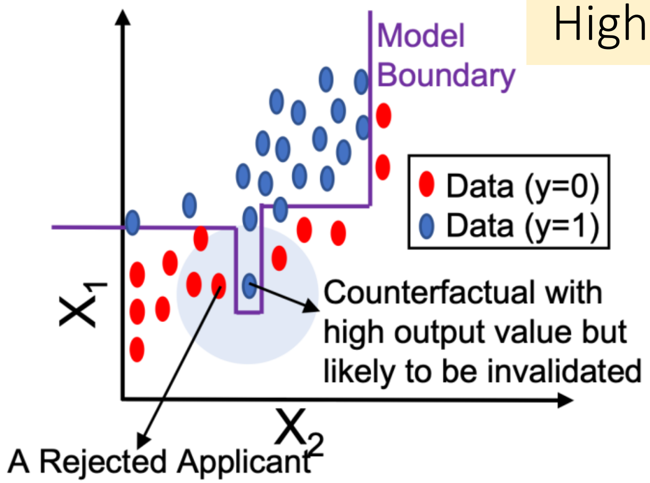
A Novel Measure to Quantify Robustness for Tree-Based Ensembles

$$R_{K,\sigma^2}(x,M) = \frac{1}{K}\sum_{x'\in N_x} M(x') - \sqrt{\frac{1}{K}\sum_{x'\in N_x}\left(M(x') - \frac{1}{K}\sum_{x'\in N_x} M(x')\right)^2}$$

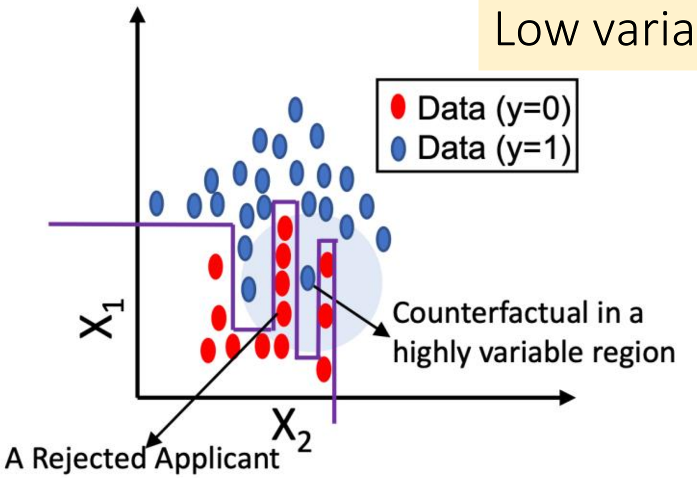
where N_x is a set of K points from the distribution $N(x,\sigma^2)$



High mean



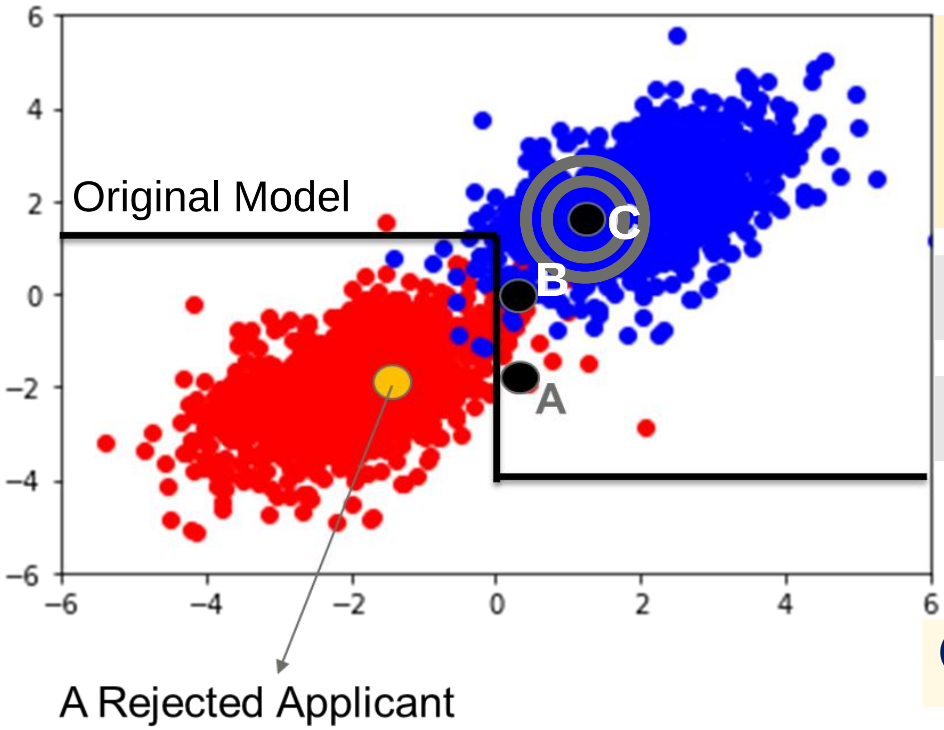
Low variance



Identify key properties that affect robustness

Contribution 2: Conservative Counterfactuals

Nearest neighbor in the dataset on the other side of the decision boundary that also has high stability, i.e., $R_{K,\sigma^2}(x, M) \geq \tau$ (stability test)



C: Conservative Counterfactual
(Closest Data-Support Counterfactual
that is also Well-Within the boundary)

B: Closest Data-Support Counterfactual

A: Closest Counterfactual

Theoretical Robustness Guarantee

Contribution 3: RobX Algorithm

Finds counterfactuals that are close, robust, and realistic

- Can be applied on top of any base-method of counterfactual generation for tree-based models, e.g., Feature Tweaking, FOCUS, FACE, kNN, etc.
- Iteratively refines the generated counterfactual and keeps moving it towards a conservative counterfactual until $R_{K,\sigma^2}(x, M) \geq \tau$ (stability test)

Experimental Results on GERMAN CREDIT and HELOC datasets:
More robust (validity) and realistic (LOF) with slight increase in distance (Lp norm)

Thank You!