TEACHPRIVACY

GDPR's Scope

www.teachprivacy.com

Written by Daniel J. Solove

Illustrated by Ryan Beckwith

Permission to use granted on: Oct 28, 2019

Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR

Anna, Emma, and Kaivalya

Written by:
Sandra Wachter (law/ethics)
Brent Mittelstadt (philosophy/ethics)
Chris Russell (machine learning/ethics)

Outline

- GDPR
- Counterfactuals
- Experiments

Why explain?

A) Explanations to UNDERSTAND decisions
B) Explanations to CONTEST decisions
C) Explanations to ALTER FUTURE decisions

GDPR Crash Course

What?

“The General Data Protection Regulation (GDPR) codifies and unifies the data privacy laws across all the EU member countries.”

https://www.techrepublic.com/article/the-eu-general-data-protection-regulation-gdpr-the-smart-persons-guide/

Who?

“The GDPR is applicable to any citizen of the European Union and, most importantly, for any company doing business with a citizen of the EU.”

https://www.techrepublic.com/article/the-eu-general-data-protection-regulation-gdpr-the-smart-persons-guide/

Why care?

“the penalties laid out for violations are significant. Enterprises found to be in violation of the provisions of the GDPR can be fined up to 4% of annual global turnover or 20 Million Euros, whichever is greater.”

($22.3m USD)

https://www.techrepublic.com/article/the-eu-general-data-protection-regulation-gdpr-the-smart-persons-guide/

Date	Organisation	Amount	Issued by	Reason(s)
2019-01-21	Google LLC	€50 million	France (CNIL)	Insufficient transparency, control, and consent over the processing of personal data for the purposes of behavioural advertising.[5][6]
2019-07-09	Marriott International	£99 million	UK (ICO)	Failure to undertake sufficient due diligence when acquiring Starwood hotels group, whose systems where compromised in 2014, exposing approximately 339 million guest records[25]
2019-07-08	British Airways	£183 million	UK (ICO)	Use of poor security arrangements that resulted in a 2018 web skimming attack affecting 500,000 consumers.[20][21][22]

https://en.wikipedia.org/wiki/GDPR_fines_and_notices

Date	Organisation	Amount	Issued by	Reason(s)
2019-09-20	Online retailer Morele.net	€645,000	Poland (UODO)	Insufficient protection of personal data, leading to the exposure of data of about 2.2 million people[39]
2019-07-16	HagaZiekenhuis	€460,000	The Netherlands (AP)	Insufficient security of medical records[28][29]
2018-10	Hospital do Barreiro	€400,000	Portugal (CNPD)	"...based on access policies to databases, which allowed technicians and physicians to consult patients' clinical files, without proper authorization."[2]
2019-06-18	Sergic (real estate services)	€400,000	France (CNIL)	Failure to implement appropriate security measures; failure to define appropriate data retention periods for the personal data of unsuccessful rental candidates.[15]

https://en.wikipedia.org/wiki/GDPR_fines_and_notices

When?

“Enforcement of the GDPR went into effect May 25, 2018.”

https://www.techrepublic.com/article/the-eu-general-data-protection-regulation-gdpr-the-smart-persons-guide/

What are the main provisions?

- Informed Consent (intelligible, clear, easy to withdraw)
- Rights:
  - Breach notifications
  - Right to access and information
  - Right of erasure, rectification
  - Data portability
  - Contest automated decisions
- Principles:
  - Data minimization
  - Security

https://eugdpr.org/the-regulation/

What *is* it?

The GDPR contains 99 articles and 173 recitals! What's the difference?

> "the GDPR consists of two components: the articles and recitals. The articles constitute the legal requirements organizations must follow to demonstrate compliance. The recitals provide additional information and supporting context to supplement the articles.
>
> The European Data Protection Board—formerly Article 29 Working Party—relies on the recitals to interpret the articles. Furthermore, the Court of Justice of the European Union reviews the recitals to decide the meaning and application of the GDPR"

https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/practice/2019/a-very-brief-introduction-to-the-gdpr-recitals/

Their two main arguments:

1) The GDPR does not require “opening the black box”
2) Counterfactual explanations fulfill (and go beyond) the requirements of the GDPR

Our two main goals:

1) Read the key articles & recitals Wachter et al. rely on
2) See what they do and do not say (show ambiguity)

Why explain?

A) **Explanations to UNDERSTAND decisions**
B) Explanations to CONTEST decisions
C) Explanations to ALTER FUTURE decisions

GDPR Article 13

1. Where personal data relating to a data subject are collected from the data subject, the controller shall, at the time when personal data are obtained, provide the data subject with all of the following information:
   a. the identity and the contact details of the controller and, where applicable, of the controller's representative;
   b. the contact details of the data protection officer, where applicable;
   c. the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;
   d. where the processing is based on point (f) of Article 6(1), the legitimate interests pursued by the controller or by a third party;
   e. the recipients or categories of recipients of the personal data, if any;
   f. where applicable, the fact that the controller intends to transfer personal data to a third country or international organisation and the existence or absence of an adequacy decision by the Commission, or in the case of transfers referred to in Article 46 or 47, or the second subparagraph of Article 49(1), reference to the appropriate or suitable safeguards and the means by which to obtain a copy of them or where they have been made available.
2. In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:
   a. the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
   b. the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject or to object to processing as well as the right to data portability;
   c. where the processing is based on point (a) of Article 6(1) or point (a) of Article 9(2), the existence of the right to withdraw consent at any time, without affecting the lawfulness of processing based on consent before its withdrawal;
   d. the right to lodge a complaint with a supervisory authority;
   e. whether the provision of personal data is a statutory or contractual requirement, or a requirement necessary to enter into a contract, as well as whether the data subject is obliged to provide the personal data and of the possible consequences of failure to provide such data;
   f. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.
3. Where the controller intends to further process the personal data for a purpose other than that for which the personal data were collected, the controller shall provide the data subject prior to that further processing with information on that other purpose and with any relevant further information as referred to in paragraph 2.
4. Paragraphs 1, 2 and 3 shall not apply where and insofar as the data subject already has the information.

GDPR Article 13 (2) f

...the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:...

the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

GDPR Article 14 (1-2)

1. Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information:
   a. the identity and the contact details of the controller and, where applicable, of the controller's representative;
   b. the contact details of the data protection officer, where applicable;
   c. the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;
   d. the categories of personal data concerned;
   e. the recipients or categories of recipients of the personal data, if any;
   f. where applicable, that the controller intends to transfer personal data to a recipient in a third country or international organisation and the existence or absence of an adequacy decision by the Commission, or in the case of transfers referred to in Article 46 or 47, or the second subparagraph of Article 49(1), reference to the appropriate or suitable safeguards and the means to obtain a copy of them or where they have been made available.
2. In addition to the information referred to in paragraph 1, the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject:
   a. the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
   b. where the processing is based on point (f) of Article 6(1), the legitimate interests pursued by the controller or by a third party;
   c. the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject and to object to processing as well as the right to data portability;
   d. where processing is based on point (a) of Article 6(1) or point (a) of Article 9(2), the existence of the right to withdraw consent at any time, without affecting the lawfulness of processing based on consent before its withdrawal;
   e. the right to lodge a complaint with a supervisory authority;
   f. from which source the personal data originate, and if applicable, whether it came from publicly accessible sources;
   g. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

GDPR Article 14 (3-5)

3. The controller shall provide the information referred to in paragraphs 1 and 2:
   a. within a reasonable period after obtaining the personal data, but at the latest within one month, having regard to the specific circumstances in which the personal data are processed;
   b. if the personal data are to be used for communication with the data subject, at the latest at the time of the first communication to that data subject; or
   c. if a disclosure to another recipient is envisaged, at the latest when the personal data are first disclosed.
4. Where the controller intends to further process the personal data for a purpose other than that for which the personal data were obtained, the controller shall provide the data subject prior to that further processing with information on that other purpose and with any relevant further information as referred to in paragraph 2.
5. Paragraphs 1 to 4 shall not apply where and insofar as:
   a. the data subject already has the information;
   b. the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available;
   c. obtaining or disclosure is expressly laid down by Union or Member State law to which the controller is subject and which provides appropriate measures to protect the data subject's legitimate interests; or
   d. where the personal data must remain confidential subject to an obligation of professional secrecy regulated by Union or Member State law, including a statutory obligation of secrecy.

GDPR Article 15

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:
    a. the purposes of the processing;
    b. the categories of personal data concerned;
    c. the recipients or categories of recipient to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;
    d. where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;
    e. the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;
    f. the right to lodge a complaint with a supervisory authority;
    g. where the personal data are not collected from the data subject, any available information as to their source;
    h. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.
2. Where personal data are transferred to a third country or to an international organisation, the data subject shall have the right to be informed of the appropriate safeguards pursuant to Article 46 relating to the transfer.
3. The controller shall provide a copy of the personal data undergoing processing. For any further copies requested by the data subject, the controller may charge a reasonable fee based on administrative costs. Where the data subject makes the request by electronic means, and unless otherwise requested by the data subject, the information shall be provided in a commonly used electronic form.
4. The right to obtain a copy referred to in paragraph 3 shall not adversely affect the rights and freedoms of others.

Why explain?

A) Explanations to UNDERSTAND decisions
B) Explanations to CONTEST decisions
C) Explanations to ALTER FUTURE decisions

GDPR Article 22

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
    a. is necessary for entering into, or performance of, a contract between the data subject and a data controller;
    b. is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
    c. is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Why explain?

A) Explanations to UNDERSTAND decisions
B) Explanations to CONTEST decisions
C) Explanations to ALTER FUTURE decisions

Key recitals

Recital 71: (Profiling)

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child.

In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

Recital 63: (Right of access)

A data subject should have the right of access to personal data which have been collected concerning him or her, and to exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. This includes the right for data subjects to have access to data concerning their health, for example the data in their medical records containing information such as diagnoses, examination results, assessments by treating physicians and any treatment or interventions provided. Every data subject should therefore have the right to know and obtain communication in particular with regard to the purposes for which the personal data are processed, where possible the period for which the personal data are processed, the recipients of the personal data, **the logic involved in any automatic personal data processing and, at least when based on profiling, the consequences of such processing**. Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data. That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject. Where the controller processes a large quantity of information concerning the data subject, the controller should be able to request that, before the information is delivered, the data subject specify the information or processing activities to which the request relates.

Their two main arguments:

1) The GDPR does not require “opening the black box”
2) Counterfactual explanations fulfill (and go beyond) the requirements of the GDPR

Counterfactuals

GDPR: Individual's Rights

The GDPR establishes the following rights for individuals:

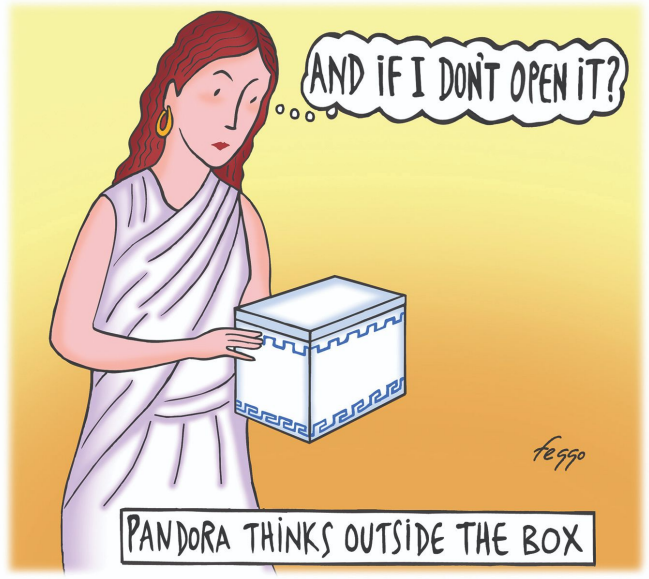
1. The right to be informed
2. The right of access
3. The right to rectification
4. The right to erasure
5. The right to restrict processing
6. The right to data portability
7. The right to object
8. Rights in relation to automated decision making and profiling⇒ RIGHT TO EXPLANATION

source:https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/

Right to Explanation

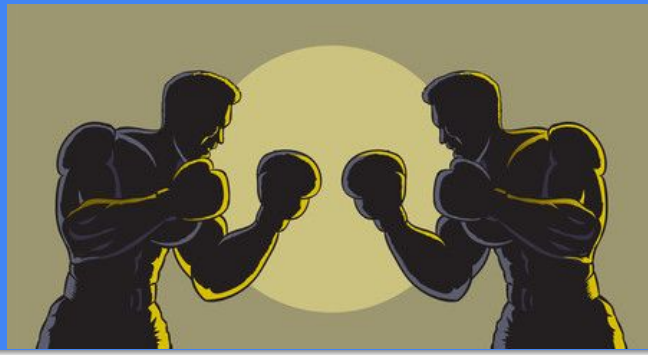
Current Approach: Open the Black Box:

- But you don't have to! Right to explanation is not legally binding.
- If it were to be biding, applies only to fully automated negative decisions with significant negative effects.
- Very difficult to actually implement and explain all aspects in practice.



Source:http://www.prospectmagazine.co.uk/wp-content/uploads/2014/12/USEDcartoon_pandora.jpg?ffd6b1

Wachter vs. Rudin: Round I



- Wachter believes that opening the black box hurts the firm as it:
  - Reveals trade secrets
  - Allows subjects to game the system.
- Rudin doesn't believe that proprietory black box models are more accurate than interpretable models and worth preserving like a trade secret.
- Rudin also doesn't believe that good models can be gamed (i.e susceptibility to being manipulated means bad design).
- Who do you think is right?

Right to Explanation

Sandra Wachter's Approach: Use Counterfactuals:

- Design explanation that helps subjects to act rather than just understand.
- Three key goals of this approach:
  - Inform and help the subject understand why a particular decision was reached.
  - Provide grounds to contest adverse decisions.
  - Understand what could be changed to receive a desired result in the future, based on the current decision-making model.
- Use unconditional ***Counterfactual Explanations*** for positive or negative decisions with significant effects delivered by fully or partially automated processes

Counterfactuals: Background

- Counterfactuals are statements in the form:
  - “You your resume was not selected because you have not taken any classes on interpretability. Had you taken CS282br, you would have been selected for the interviews.”
  - “The cat moved into the dog’s bed after the dog ate all the cat’s Fancy Feast. Had the dog stayed away from the cat’s food, the dog would would be sleeping in the doggy bed instead of the cat.”

“Following the interview, your résumé will be put through our rigorous selection process.”

source:https://gravitational.com/blog/coding-challenge/

Counterfactual: Important Distinction

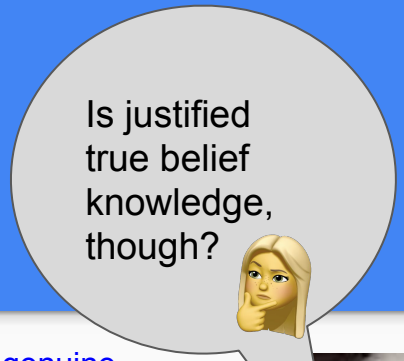
- ML and Lawyers' Approach to Explanation: Focuses primarily on explanation of the internal structure of the algorithms and how it led to the decisions.
- Counterfactual Approach to Explanation: Describes dependency on the external facts that led to the decision (i.e. works outside the black box).

Counterfactuals: Philosophical Diversion

1. Defining propositional knowledge:
   a. Statement of the form: "*S knows that p*"
   b. S is the knowing party and p is the known proposition.
2. Traditional analytical philosophy ascertains that propositional knowledge = justified true belief. Thus, the following conditions must be met of a given claim for us to have knowledge of that claim:
   a. Truth
   b. Belief
   c. Justification

Individually necessary and jointly sufficient for propositional knowledge

Philosophical Diversion



Gettier's Counterexamples are situations where three conditions are satisfied but no genuine knowledge is achieved:

1) One can believe in something that is justifiably true without actually knowing that that proposition is true.
2) One can justifiably believe case 1 or case 2 that are not related, where he has justification for just case 1. If case 1 turns out to be false but case 2 happens to be true, then the belief was true and justified but not knowledge.

Edmund L. Gettier III

3. Recent, additional condition: Sensitivity

   a. "If p were false, S would not believe that p."
   b. "If p were false" is a counterfactual that lives in a possible world where p is true.
   c. In the nearest possible world worlds in which no-p, the subject does not believe that p.

Counterfactual: Definition

Wachter extends the concept of sensitivity to update it in the following form:

"If q were false, S would not believe p."

where q serves as the explanation of S's belief in p. Now, these statements now only apply to S's belief in p.

Counterfactual Explanations are statements taking the form:

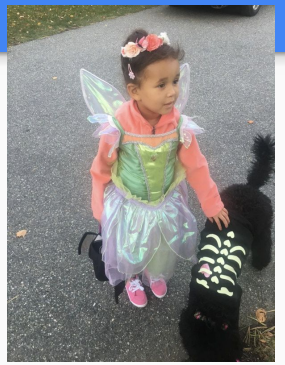
Score p was returned because variables V had values (v1, v2,...) associated with them. If V instead had values (v1', v2',...), and all other variables had remained constant, score p' would have been returned.

Counterfactual: Key Aspects

- “Closest possible world”: ideal counterfactual explanation would alter values as little as possible and represent a closest world under which score p' is returned instead of p. But this is case specific (more on this later).
- Structural equations approach in execution by identifying alterations to variables. Similar to Pearl’s “mini-surgeries” (minimal change to model to incorporate a condition) than Lewis’ “miracles.” (big change outside of the closest world - something more than nature, a supernatural world that intervenes).
- This approach does not rely on knowledge of the causal structure of the world, or suggest which context-dependent metric of distance between worlds is preferable to establish causality.

Different Approaches to Explanation

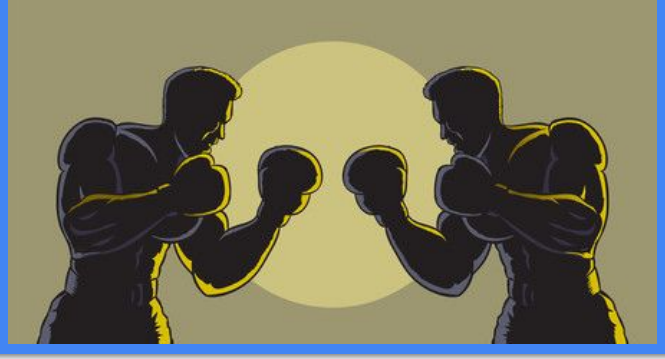
- Early work on rule based explanations similar to counterfactuals.
  - Q:Why is a tax cut appropriate?
  - A: Because a tax cut's preconditions are high inflation and trade deficits, and current conditions include these factors.
  - What would be the counterfactual version of this statement?
  - RULE#1 IF: 1) If it is Halloween and 2) there are children under 8 in the household THEN: There is strongly suggestive evidence (.8) that the the family dog(s) will be wearing a Halloween costume.
  - What would be the counterfactual version of this statement?
- What makes counterfactuals more effective?



Explanations of Machine Learning

- Wachter argues that modern explanations focus on building simplified models to explain complex algorithm.
- She argues that even with this approach they are not interpretable to lay people.
- Wachter points to a three-way trade-off between the quality of the approximation, the ease of understanding the function, and the size of the domain for which the approximation is valid.
- What do you think about this statement?
- Are counterfactuals better than the current approach to interpretability?

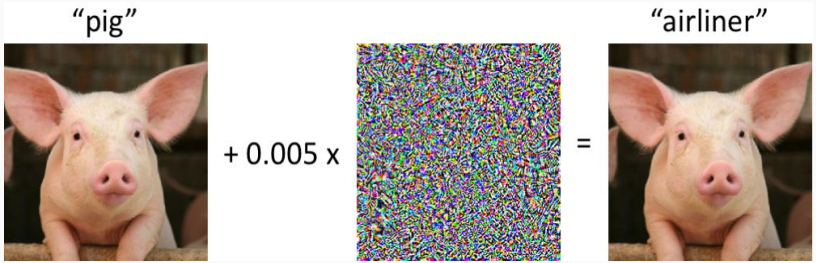
Wachter vs. Rudin: Round II

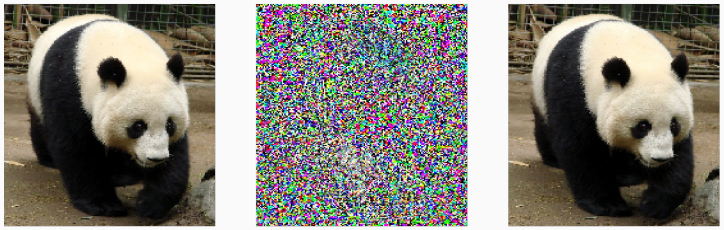


- Rudin believes that counterfactual explanations are inadequate for high stakes decisions.
  - Counterfactuals do not produce lowest cost decisions for the user.
    - Consider the following counterfactuals for the same problem:
      - “Had you reduced your debt by $5000 and increased your savings by 50% then you would have qualified for the loan you applied for”
      - “If you had gotten a job that pays $500 more per week, then you would have qualified for the loan”
  - AND require cost information from the user that is difficult to obtain.
- Wachter counters with: *“It is more informative to provide a diverse set of counterfactual explanations, corresponding to different choices of nearby possible worlds rather than a theoretically ideal counterfactual describing the “closest possible world” according to a preferred distance metric”.*

Adversarial Perturbations & Counterfactuals

- Counterfactuals are used to confuse existing classifiers by generating a synthetic data point close to an existing one such that the new synthetic data point is classified differently than the original one.
- These classifiers have been shown to be particularly vulnerable to a type of attack referred to as “Adversarial Perturbation” where small changes to a given image can result in the image being assigned to an entirely different class.





original image
prediction: giant_panda

the perturbation,
enhanced 127 times

perturbed image
prediction: bucket

Source: http://gradientscience.org/intro_adversarial/, https://www.arxiv-vanity.com/papers/1804.05296/

Adversarial Perturbations & Counterfactuals

- BUT, none of the standard works on Adversarial Perturbations make use of appropriate distance functions, and most favor making small changes to many variables, rather than modifying only a few variables.
- One of the more challenging aspects of Adversarial Perturbations is that these small perturbations of an image are barely visible, but result in very different classifier responses - often newly generated images do not lie in the “space of real-images,” but slightly outside it.

Causality and Fairness

- Counterfactuals can provide evidence that an algorithmic decision is affected by a protected variable (e.g. race), and that it may therefore be discriminatory.
- If the counterfactuals found to change someone's race, then the treatment of that individual is dependent on race BUT the converse statement is not true.
- Counterfactuals which do not modify a protected attribute cannot be used as evidence that the attribute was irrelevant to the decision (more on this later).
- Counterfactuals describe only some of the dependencies between a particular decision and specific external facts (i.e do not represent causality).

Wachter’s Three Goals Reviewed

1) Inform and help the subject understand why a particular decision was reached.
2) Provide grounds to contest adverse decisions.
3) Understand what could be changed to receive a desired result in the future, based on the current decision-making model.

Experiments
(Generating Counterfactuals)

Machine Learning

Machine Learning = Optimisation?

Machine Learning

Machine Learning = **Gradient Based** Optimisation?

Machine Learning

Machine Learning = **Gradient Based** Optimisation?

Counterexamples?

Machine Learning

"Many standard classifiers are trained by finding the optimal set of weights that minimises an objective"

$$\arg\min_{w} \ell(f_w(x_i), y_i) + \rho(w)$$

Machine Learning

"Many standard classifiers are trained by finding the optimal set of weights that minimises an objective"

$$\arg\min_{w} \ell(f_w(x_i), y_i) + \rho(w)$$

$\ell(y_1, y_2) = ?$

$\rho(w) = ?$

Counterfactual Generation

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

Counterfactual Generation

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

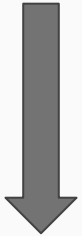
?

Counterfactual Generation: Objective

$$f_w(\mathbf{x}_i) = \mathbf{y}_i$$

Counterfactual Generation: Objective

$$f_w(x_i) = y_i$$



$$f_w(x') = y' \quad (1)$$

$$x' \cong x_i \quad (2)$$

Counterfactual Generation: Objective (1)

$$f_w(\mathbf{x'}) = \mathbf{y'}$$

Counterfactual Generation: Objective (1)

$$f_w(\mathbf{x'}) = \mathbf{y'}$$

$$f_w(\mathbf{x'}) \cong \mathbf{y'}$$

Counterfactual Generation: Objective (1)

$$f_w(\mathbf{x}') = \mathbf{y}'$$

$$f_w(\mathbf{x}') \cong \mathbf{y}'$$

$$\min \ell(f_w(\mathbf{x}'), \mathbf{y}')$$

$$\ell(f_w(\mathbf{x}'), \mathbf{y}') \leq \varepsilon$$

Counterfactual Generation: Objective (1)

$$\ell(f_w(\mathbf{x'}), \mathbf{y'}) = (f_w(\mathbf{x'}) - \mathbf{y'})^2$$

Counterfactual Generation: Objective (2)

$$x' \cong x_i$$

Counterfactual Generation: Objective (2)

$$\mathbf{x}' \cong \mathbf{x}_i$$

$$\ell(\mathbf{x}', \mathbf{x}_i) \leq \varepsilon$$

$$\min d(\mathbf{x}', \mathbf{x}_i)$$

Counterfactual Generation: Objective (2)

$$\mathbf{x}' \cong \mathbf{x}_i$$

$$\ell(\mathbf{x}', \mathbf{x}_i) \leq \varepsilon$$

$$\min d(\mathbf{x}', \mathbf{x}_i)$$

$$d(\mathbf{x}', \mathbf{x}_i) = (\mathbf{x}' - \mathbf{x}_i)^2 / \sigma_x$$

Counterfactual Generation: Objective (2)

$$d(x_i, x') = \sum_{k \in F} \frac{|x_{i,k} - x'_k|}{\text{MAD}_k}$$

Counterfactual Generation: Objective

$$\lambda(f_w(x') - y')^2 + d(x_i, x')$$

Counterfactual Generation: Objective

$$\lambda(f_w(x') - y')^2 + d(x_i, x')$$

(minimise)

Counterfactual Generation: Objective

$$\lambda(f_w(x') - y')^2 + d(x_i, x')$$

(minimise)

$$(f_w(\mathbf{x'}) - \mathbf{y'})^2 \leq \varepsilon$$

Counterfactual Generation: Objective

$$\lambda(f_w(x') - y')^2 + d(x_i, x')$$

(minimise)

$$(f_w(\mathbf{x'}) - \mathbf{y'})^2 \leq \varepsilon$$

Counterfactual Generation: LSAT

Original Data			
Score	GPA	LSAT	Race
0.17	3.1	39.0	0
0.54	3.7	48.0	0
-0.77	3.3	28.0	1
-0.83	2.4	28.5	1
-0.57	2.7	18.3	0

Counterfactual Generation: LSAT

Counterfactuals			
Score	GPA	LSAT	Race
0.0	(- 0.1) 3.0	39.0	0.3
0.0	(- 0.2) 3.5	(- 0.1) 47.9	0.9
0.0	(+ 0.2) 3.5	(+ 0.1) 28.1	-0.3
0.0	(+ 0.2) 2.6	(+ 0.1) 28.6	-0.4
0.0	(+ 0.2) 2.9	(+ 0.1) 18.4	-1.0

Counterfactual Generation: LSAT

Counterfactuals			
Score	GPA	LSAT	Race
0.0	(- 0.1) 3.0	39.0	0.3
0.0	(- 0.2) 3.5	(- 0.1) 47.9	0.9
0.0	(+ 0.2) 3.5	(+ 0.1) 28.1	-0.3
0.0	(+ 0.2) 2.6	(+ 0.1) 28.6	-0.4
0.0	(+ 0.2) 2.9	(+ 0.1) 18.4	-1.0

Counterfactual Generation: LSAT

Counterfactual Hybrid			
Score	GPA	LSAT	Race
0.0	(- 1.6) 1.5	(- 0.6) 38.4	0
0.0	(- 5.3) -1.6	(- 2.1) 45.9	0
0.0	(+ 2.0) 5.3	(+ 0.9) 28.9	0
0.0	(+ 2.4) 4.8	(+ 0.9) 29.4	0
0.0	(+ 5.7) 8.4	(+ 2.3) 20.6	0

Counterfactual Generation: LSAT

Counterfactual Hybrid			
Score	GPA	LSAT	Race
0.0	(- 1.6) 1.5	(- 0.6) 38.4	0
0.0	(- 5.3) -1.6	(- 2.1) 45.9	0
0.0	(+ 2.0) 5.3	(+ 0.9) 28.9	0
0.0	(+ 2.4) 4.8	(+ 0.9) 29.4	0
0.0	(+ 5.7) 8.4	(+ 2.3) 20.6	0

Counterfactual Generation: LSAT

Counterfactual Hybrid (L2 Normalised)			
Score	GPA	LSAT	Race
0.0	(- 0.1) 3.0	(- 5.0) 34.0	0
0.0	(- 0.2) 3.5	(- 14.9) 33.1	0
0.0	(+ 0.1) 3.4	(+ 5.4) 33.4	0
0.0	(+ 0.2) 2.6	(+ 7.2) 35.7	0
0.0	(+ 0.2) 2.9	(+ 15.8) 34.1	0

Counterfactual Generation: LSAT

Counterfactual Hybrid (L2 Normalised)			
Score	GPA	LSAT	Race
0.0	(- 0.1) 3.0	(- 5.0) 34.0	0
0.0	(- 0.2) 3.5	(- 14.9) 33.1	0
0.0	(+ 0.1) 3.4	(+ 5.4) 33.4	0
0.0	(+ 0.2) 2.6	(+ 7.2) 35.7	0
0.0	(+ 0.2) 2.9	(+ 15.8) 34.1	0

Counterfactual Generation: LSAT

Counterfactual Hybrid (L1 Normalised)			
Score	GPA	LSAT	Race
0.0	3.1	(- 5.0) 34.0	0
0.0	3.7	(- 15.6) 32.4	0
0.0	3.3	(+ 5.5) 33.5	0
0.0	2.4	(+ 7.3) 35.8	0
0.0	2.7	(+ 16.6) 34.9	0

Counterfactual Generation: Pima

Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

“What would have to be different for this individual to have a risk score of 0.5?”

Counterfactual Generation: Pima

Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

“What would have to be different for this individual to have a risk score of 0.5?”

- Person 1: If your 2-Hour serum insulin level was 154.3, you would have a score of 0.51.
- Person 2: If your 2-Hour serum insulin level was 169.5, you would have a score of 0.51.
- Person 3: If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5, you would have a score of 0.51.

Questions ?

Actionable Recourse in Linear Classification

Ustun, Spanger, Liu, 2018

Slides and Presentation by Jason Ren

Agenda

1. Background and Terminology
2. Contributions
3. Problem Statement
   - Optimization framework, feasibility and cost guarantees
4. Methods
   - IP Formulation, cost functions, constructing flipsets
5. Results
6. Discussion

Background and Terminology

- Recourse: ability of a person to obtain a desired outcome from a fixed prediction model
- Actionable vs Immutable vs Conditionally Immutable Features
- Adjusted Framework: model fixed, data variable

Contributions

Ustun et. al. introduce:

1. A procedure to measure the feasibility and cost of recourse of a linear classifier for individuals in a target population.
2. A procedure to generate a list of optimal actions that an individual can make to flip the prediction of the classifier (flipset).

Related Work and Differences:

1. Application of inverse classification
2. Related to methods that explain predictions of ML models at individual level, but provides means to analyze feasibility and cost of recourse.
3. Builds on Wachter et. al.'s counterfactual explanation work to provide feasibility and optimality guarantees to audit recourse

Problem Statement

Given a linear classifier $f(\mathbf{x}; \mathbf{w}) = sign(\mathbf{w}^\top \mathbf{x}) : \mathbb{R}^{d+1} \rightarrow \{-1, 1\}$,

For any individual $\mathbf{x}^*$ s.t. $f(\mathbf{x}^*) = -1$,

Determine does there exist (and if there does what is the optimal) an action $\mathbf{a}$ s.t. $f(\mathbf{x}^* + \mathbf{a}) = 1$

Optimization Framework

Can formalize the problem statement as:

Given an individual $\mathbf{x}$ s.t. $f(\mathbf{x}) = -1$:

$$\begin{aligned} &\min \text{ cost } (\mathbf{a}, \mathbf{x}) \\ &\text{s.t. } f(\mathbf{x} + \mathbf{a}) = 1 \\ &\mathbf{a} \in A(\mathbf{x}) \end{aligned}$$

where:

- $A(\mathbf{x})$ is the set of feasible actions from $\mathbf{x}$.
- $\text{cost}(\cdot, \mathbf{x}) : A(\mathbf{x}) \to \mathbb{R}_+$ is a user-specified cost function

Feasibility Guarantees

Remark 1: A linear classifier provides recourse to all individuals $\mathbf{x} \in H^-$ if all features are actionable and it does not trivially predict a single class.

Remark 2: If the feature values belong to a unbounded* space, then a linear classifier w/ at least one actionable feature provides recourse to all individuals in any target population.

Remark 3: If the feature values belong to a bounded space, then a linear classifier w/ at least one immutable feature may not provide recourse to some individuals in a target population.

Cost Guarantees

$$E_{H^-}\left[cost(\mathbf{a}_{opt};\mathbf{x})\right] \leq \frac{1}{||\mathbf{w}_A||_2^2}\left(\pi c_{D^+} - (1-\pi)c_{D^-} + 2c_{max}R_A(f)\right)$$

where:

- $\text{cost}(\mathbf{a};\mathbf{x}) = c(\mathbf{x}) \cdot ||\mathbf{a}||$
- $\mathbf{w}_A$ are the coefficients for the actionable features
- $\pi = P_{H^-}(y = 1)$ is the false omission rate of f
- $c_{D^+} = E_{H^- \cap D^+}\left[c(\mathbf{x}) \cdot \mathbf{w}_A^\top \mathbf{x}_A\right]$ is a bound on the expected cost of recourse for false negatives
- $c_{D^-} = E_{H^- \cap D^-}\left[c(\mathbf{x}) \cdot \mathbf{w}_A^\top \mathbf{x}_A\right]$
- $c_{\text{max}} = \max_{x\ \in H^-} |c(\mathbf{x}) \cdot \mathbf{w}_A^\top \mathbf{x}_A|$
- $R_A(f) = \pi \cdot P_{H^- \cap D^+}(\mathbf{w}_A^\top \mathbf{x}_A \leq 0) + (1-\pi) P_{H^- \cap D^-}(\mathbf{w}_A^\top \mathbf{x}_A \geq 0)$ is the internal risk of $\mathbf{w}_A$ for $\mathbf{x} \in H^-$.

Cost Guarantee Takeaways

$$E\left[cost(\mathbf{a}_{opt};\mathbf{x})\right] \leq \frac{P(f(\mathbf{x}) = -1)}{||\mathbf{w}_A||_2^2}\left(\pi c_{D^+} - (1-\pi)c_{D^-} + 2c_{max}R_A(f)\right)$$

We can reduce expected cost of recourse by reducing

- The maximum cost of recourse c_{max}
- The internal risk $R_A(f)$
- $P(f(\mathbf{x}) = -1)$

IP Formulation

There exists an IP Formulation of

$$\begin{aligned} &\min \text{ cost } (\mathbf{a}, \mathbf{x}) \\ &\text{s.t. } f(\mathbf{x} + \mathbf{a}) = 1 \\ &\mathbf{a} \in A(\mathbf{x}) \end{aligned}$$

with no discretization error for feasibility of recourse and discretization error for cost that can be controlled by refining the discrete grid.

IP Formulation

$$\min \quad \text{cost}$$

$$\text{s.t.} \quad \text{cost} = \sum_{j \in J_A} \sum_{k=1}^{m_j} c_{jk} v_{jk} \tag{2a}$$

$$\sum_{j \in J_A} w_j a_j \geq \sum_{j=0}^{d} w_j x_j \tag{2b}$$

$$a_j = \sum_{k=1}^{m_j} a_{jk} v_{jk} \qquad j \in J_A \tag{2c}$$

$$1 = u_j + \sum_{k=1}^{m_j} v_{jk} \qquad j \in J_A \tag{2d}$$

$$a_j \in \mathbb{R} \qquad j \in J_A$$

$$u_j \in \{0, 1\} \qquad j \in J_A$$

$$v_{jk} \in \{0, 1\} \qquad k = 1 \ldots m_j, \; j \in J_A$$

Cost Function Notes and Suggestions

Should be used to encode preferences between feasible actions, should *not* be used to penalize infeasibile actions

Some options:

$$cost(\mathbf{a}, \mathbf{x}) = \max_{j \in J_A} |F_j(x_j + a_j) - F_j(x_j)|$$

$$cost(\mathbf{a}, \mathbf{x}) = \sum_{j \in J_A} \log \left(\frac{1 - F_j(x_j + a_j)}{1 - F_j(x_j)} \right)$$

Building Flipsets

Algorithm 1 Enumerate T Minimal Cost Actions for Flipset

Input

IP — instance of (2) for coefficients $\boldsymbol{w}$, features $\boldsymbol{x}$, and actions $A(\boldsymbol{x})$

$T \geq 1$ — number of items in flipset

Initialize

$\mathcal{A} \leftarrow \emptyset$ — actions shown in flipset

1: **repeat**

2: $\boldsymbol{a}^* \leftarrow$ optimal solution to IP

3: $\mathcal{A} \leftarrow \mathcal{A} \cup \{\boldsymbol{a}^*\}$ — *add $\boldsymbol{a}*$ to set of optimal actions*

4: $S \leftarrow \{j : a_j^* \neq 0\}$ — *indices of features altered by $\boldsymbol{a}*$*

5: add constraint to IP to remove actions that alter features $j \in S$:

$$\sum_{j \in S}(1 - u_j) + \sum_{j \in S} u_j \leq d - 1.$$

6: **until** $|\mathcal{A}| = T$ or IP is infeasible

Output: $\mathcal{A}$ — actions shown in flipset

Experiments

Context: Audit recourse on linear classifiers for credit scoring

Goals:

1. Show how this procedure can provide useful information for different stakeholders (i.e. credit seekers, credit providers, policy-makers)
2. Show how the feasibility and cost of recourse can be affected by common modeling practices.

Experiment 1: Model Selection

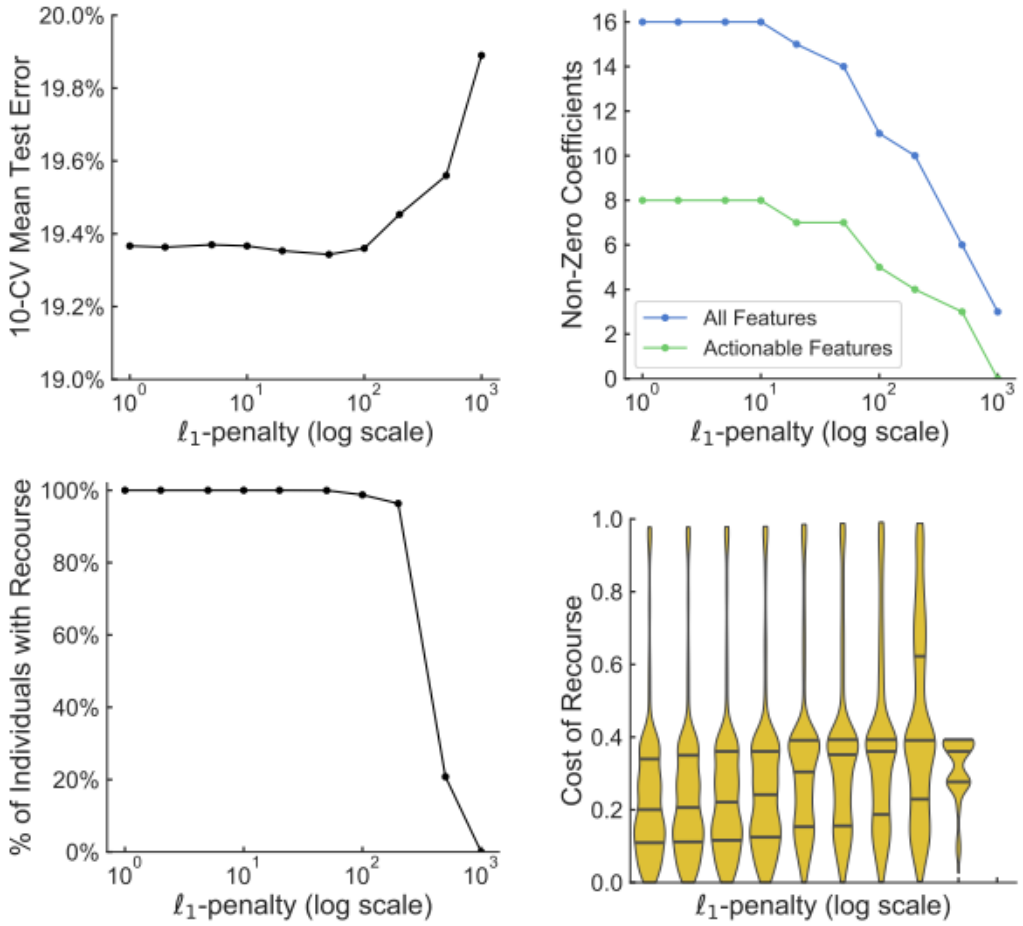


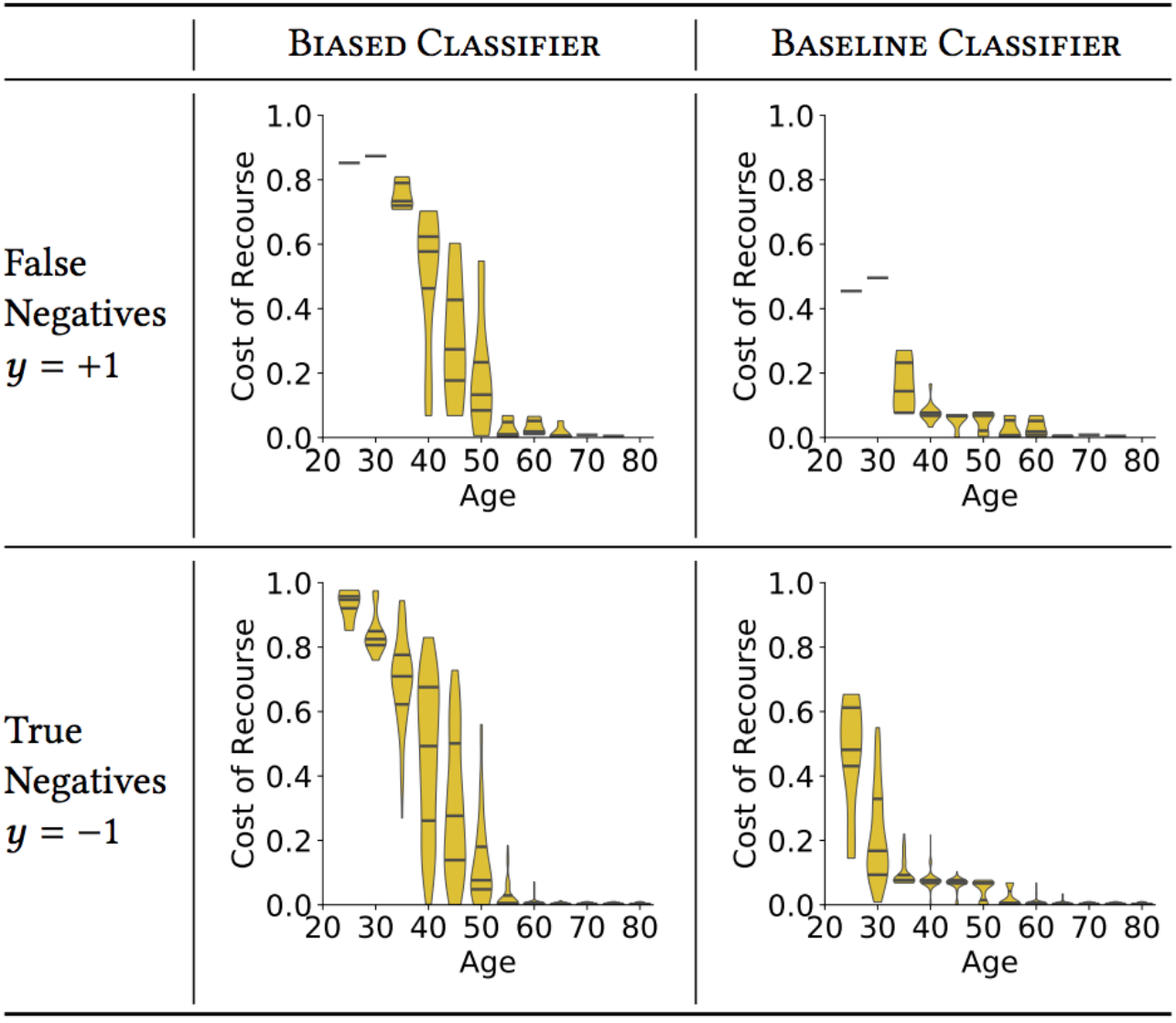
Figure 2: Model performance and recourse over the training sample for ℓ_1-penalized LR classifiers. We show the mean 10-CV test error (top left), # of non-zero coefficients (top right), % of individuals with recourse (bottom left), and the distribution of the cost of recourse (bottom right) for all classifiers.

Experiment 1: Flipset

Feature Subset	Current Values		Required Values
MostRecentPaymentAmount	\$0	$\longrightarrow$	\$790
MostRecentPaymentAmount	\$0	$\longrightarrow$	\$515
MonthsWithZeroBalanceOverLast6Months	1	$\longrightarrow$	2
MonthsWithZeroBalanceOverLast6Months	1	$\longrightarrow$	4
MostRecentPaymentAmount	\$0	$\longrightarrow$	\$775
MonthsWithLowSpendingOverLast6Months	6	$\longrightarrow$	5
MostRecentPaymentAmount	\$0	$\longrightarrow$	\$500
MonthsWithLowSpendingOverLast6Months	6	$\longrightarrow$	5
MonthsWithZeroBalanceOverLast6Months	1	$\longrightarrow$	2

Figure 3: Flipset for a person who is denied credit by the most accurate classifier built for credit. Each item describes minimal-cost changes for the individual to attain the desired outcome. We enumerated all 5 items in ≤ 1 second using the cost function in (4) and Algorithm 1.

Experiment 2: Out of Sample deployment



Experiment 2: Flipset

Biased Classifier

Feature	Current Value		Required Value
NumberOfTime60-89DaysPastDueNotWorse	0	⟶	2
NumberOfOpenCreditLinesAndLoans	1	⟶	12
MonthlyIncome	$3416	⟶	$23000
DebtRatio	0.006731	⟶	2439.73

Baseline Classifier

Feature	Current Value		Required Value
NumberOfTime60-89DaysPastDueNotWorse	0	⟶	2
NumberOfOpenCreditLinesAndLoans	1	⟶	5
MonthlyIncome	$3416	⟶	$5980
DebtRatio	0.006731	⟶	49.73

Figure 5: Minimal cost actions for an individual in the baseline population with *Age* = 32. We show actions that will result in approval from the biased classifier (top) and the baseline classifier (bottom).

Experiment 3: Evaluating Disparities in Recourse

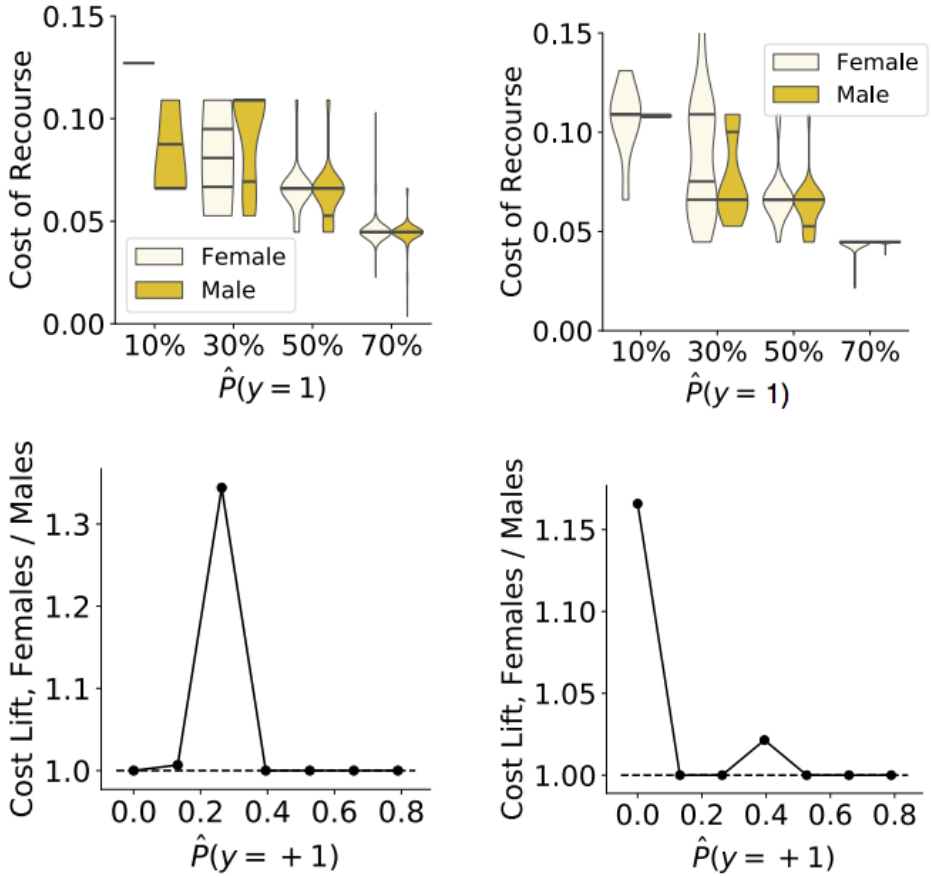


Figure 6: Overview of recourse disparities between males and females in the target population. On the top row, we plot the distribution of the cost of recourse for males and females based on their predicted risk and their true label: we plot the cost for individuals where $y = +1$ (left) and $y = -1$ (right). In the bottom row, we show the ratio of the median cost of recourse for females/males for individuals where $y = +1$ (left) and $y = -1$ (right).

Discussion

Paper strengths? Paper weaknessses?

Discussion cont.

"Recourse should be treated as an independent policy objective because it":

1. reflects a fundamnetal notion of justice (i.e. that individuals should have meaningful agency over decisions that impact their livelihood)
2. it is a precisely defined notion that can be practically regulated in many real-world applications.

What about

- Trade offs between recourse and predictive accuracy
- Manipulation Concerns