# Model Validation and Fairness in Machine Learning

2024 Marcus Evans – Model Validation Class

Nicholas Schmidt, SolasAI

June 13, 2024

SOLAS

# Introduction

- Background: Can machine learning discriminate? What causes model discrimination?
- Discrimination and the law: Relevant statutes and the burden-shifting test
- Impact, validity, and bias: Understanding the source of the problem
- Fixing the problem: searching for less discriminatory alternative models

*NOTE: The author is not a lawyer; this presentation does not represent legal or compliance advice.*

# A Little Bit About Me

- **Nicholas Schmidt**
  - Nearly 25 years of experience applying concepts from statistics and economics to law and regulatory compliance questions.

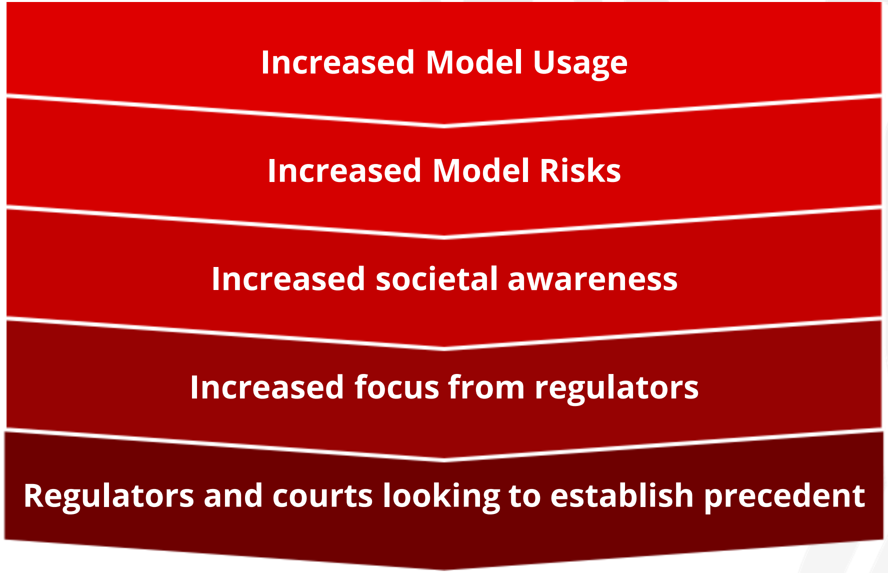
- **Founder & Chief Technology and Innovation Officer, SolasAI**
  - SolasAI software *measures* and *mitigates* discrimination risk.
  - Prominent U.S. lenders, insurers, and health insurance companies use SolasAI to assess and mitigate discrimination risk.

- **AI Practice Leader, BLDS, LLC**
  - We are the fair lending analytics advisors to lenders representing over 70% of credit cards issued in the United States.
  - Regulators and courts regularly engage us to provide guidance on discrimination risk in algorithms.

# Multiple Forces are Requiring Enterprises to Test and Justify Model Fairness

Expectations are rapidly shifting towards the need to test and justify model fairness… regardless of size or industry.  Model builders are going to be expected to:

- ✓ Test for disparities
- ✓ Identify opportunities to reduce disparity
- ✓ Provide clarity on trade-off decisions between business value and reducing disparities



Increased Model Usage

Increased Model Risks

Increased societal awareness

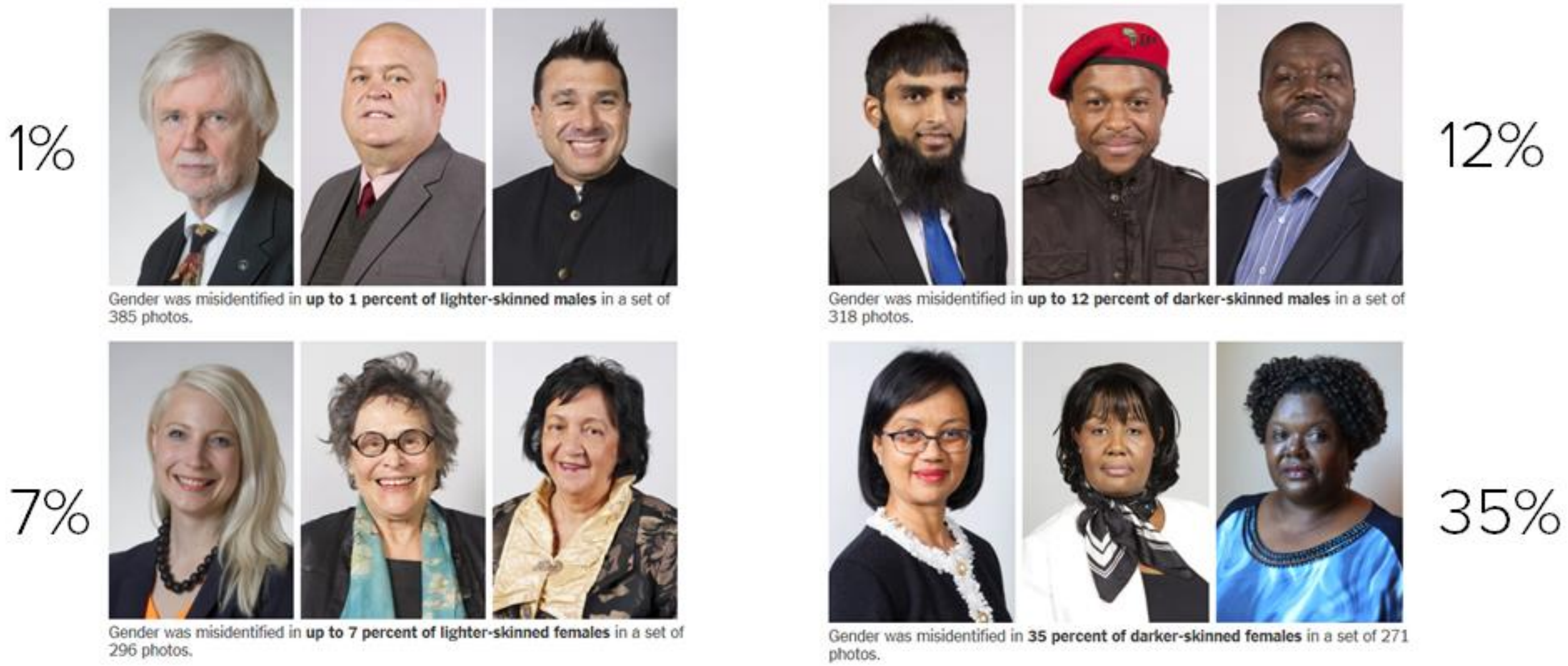
Increased focus from regulators

Regulators and courts looking to establish precedent

# Background

Can AI Discriminate?

# Can Algorithms Discriminate?



1%

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

12%

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

7%

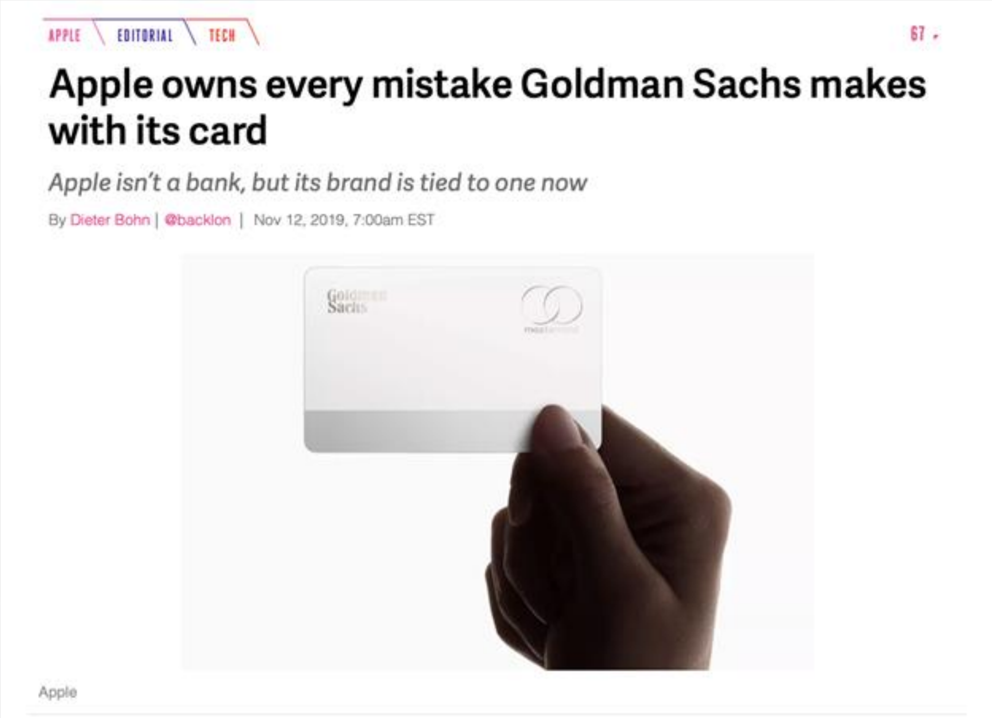
Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

35%

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Lohr, Steve. "Facial recognition is accurate, if you're a white guy." *New York Times*, 9 February 2018.

# The Effect of Two People on Twitter…



APPLE EDITORIAL TECH

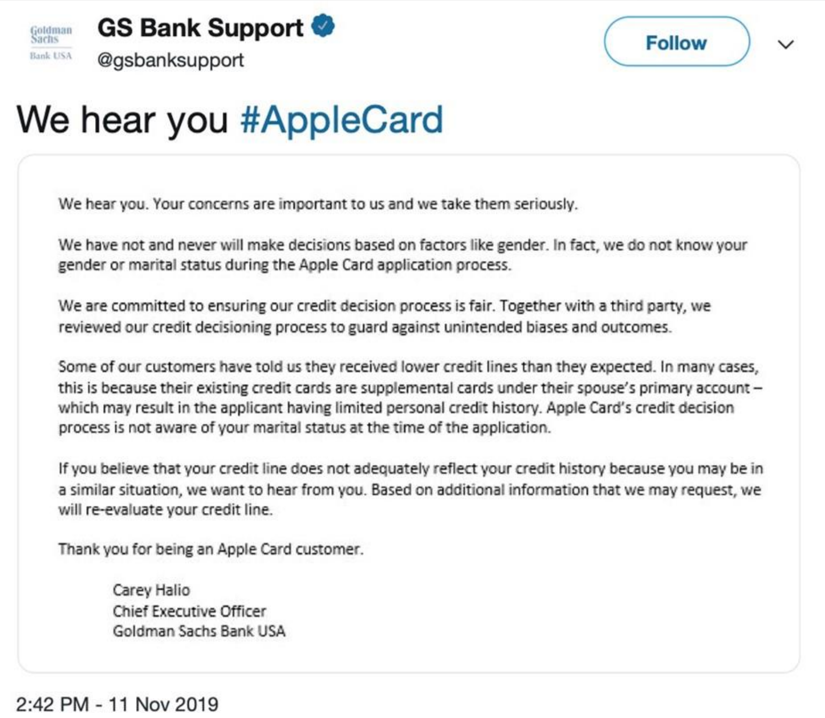
67

## Apple owns every mistake Goldman Sachs makes with its card

*Apple isn't a bank, but its brand is tied to one now*

By Dieter Bohn | @backlon | Nov 12, 2019, 7:00am EST

Apple



**GS Bank Support**
@gsbanksupport

We hear you #AppleCard

We hear you. Your concerns are important to us and we take them seriously.

We have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status during the Apple Card application process.

We are committed to ensuring our credit decision process is fair. Together with a third party, we reviewed our credit decisioning process to guard against unintended biases and outcomes.

Some of our customers have told us they received lower credit lines than they expected. In many cases, this is because their existing credit cards are supplemental cards under their spouse's primary account – which may result in the applicant having limited personal credit history. Apple Card's credit decision process is not aware of your marital status at the time of the application.

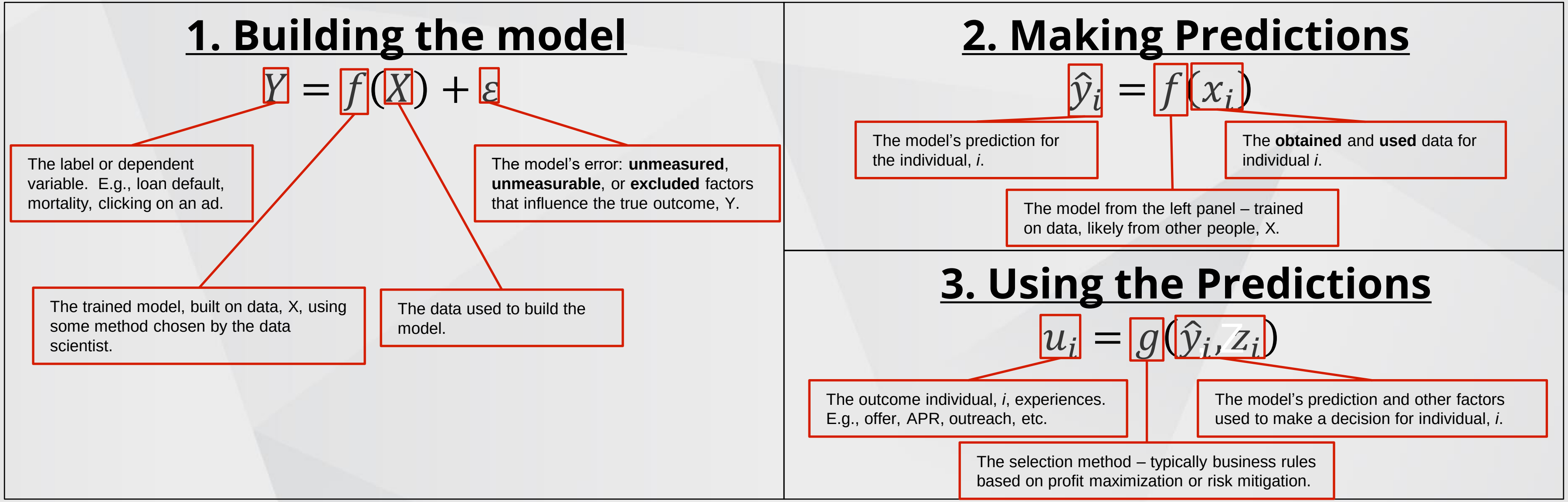
If you believe that your credit line does not adequately reflect your credit history because you may be in a similar situation, we want to hear from you. Based on additional information that we may request, we will re-evaluate your credit line.

Thank you for being an Apple Card customer.

Carey Halio
Chief Executive Officer
Goldman Sachs Bank USA

2:42 PM - 11 Nov 2019

# Where can a model discriminate?



## 1. Building the model

$$Y = f(X) + \varepsilon$$

The label or dependent variable. E.g., loan default, mortality, clicking on an ad.

The model's error: **unmeasured**, **unmeasurable**, or **excluded** factors that influence the true outcome, Y.

The trained model, built on data, X, using some method chosen by the data scientist.

The data used to build the model.

## 2. Making Predictions

$$\hat{y}_i = f(x_i)$$

The model's prediction for the individual, *i*.

The **obtained** and **used** data for individual *i*.

The model from the left panel – trained on data, likely from other people, X.

## 3. Using the Predictions

$$u_i = g(\hat{y}_i, z_i)$$

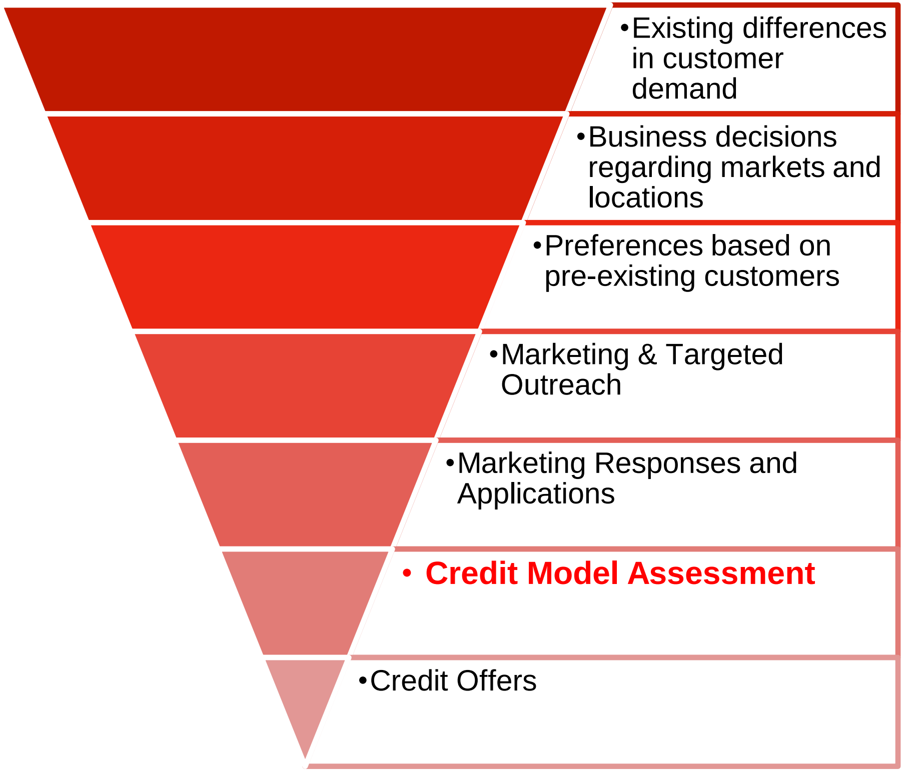
The outcome individual, *i*, experiences. E.g., offer, APR, outreach, etc.

The model's prediction and other factors used to make a decision for individual, *i*.

The selection method – typically business rules based on profit maximization or risk mitigation.

# The Problem of Model Systems

- Addressing algorithmic discrimination is essential.
- But algorithms may play a small role in a model system.
- Model systems can and should be evaluated for the potential to mitigate discrimination.



- Existing differences in customer demand
- Business decisions regarding markets and locations
- Preferences based on pre-existing customers
- Marketing & Targeted Outreach
- Marketing Responses and Applications
- **Credit Model Assessment**
- Credit Offers

# Discrimination and the Law

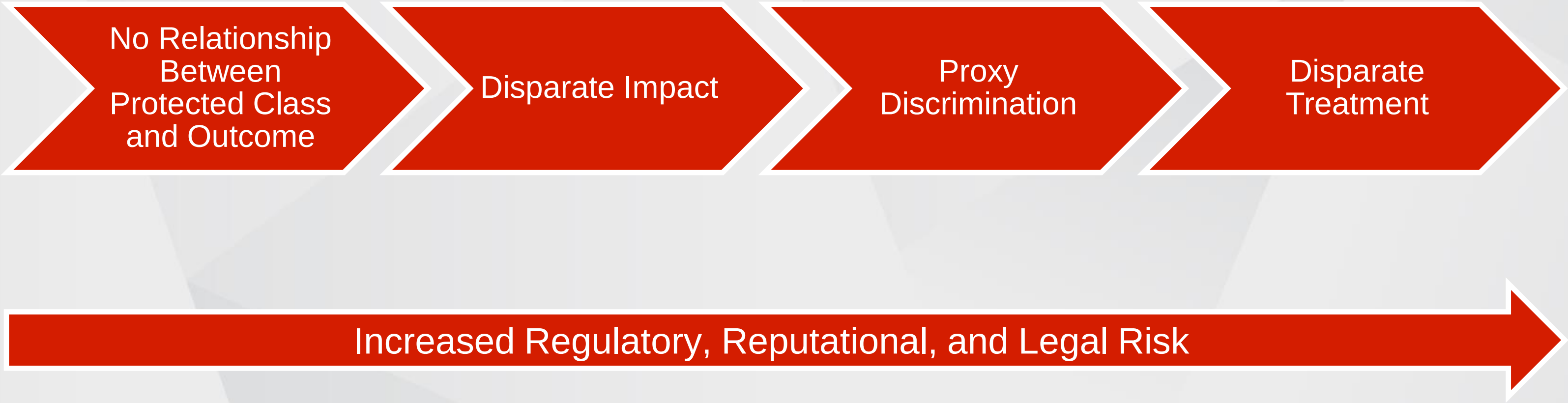
Relevant Statutes and Concepts

# Selected Relevant Federal Statutes

- Equal Credit Opportunity Act (ECOA): Prohibits discrimination in consumer credit.
- Fair Housing Act (FHA): Prohibits discrimination in most housing-related transactions.
- Title VII of the Civil Rights Act of 1964: Prohibits discrimination in employment.

| Law | Primary Protected Classes |
|---|---|
| Equal Credit Opportunity Act (ECOA) | Race, Color, Religion, National Origin, Sex, Marital Status, Age, Receipt of Public Assistance |
| Title VII of the Civil Rights Act | Race, Color, Religion, National Origin, Sex |
| Fair Housing Act (FHA) | Race, Color, Religion, National Origin, Sex, Familial Status, Disability |

# Types of Discrimination



No Relationship Between Protected Class and Outcome

Disparate Impact

Proxy Discrimination

Disparate Treatment

Increased Regulatory, Reputational, and Legal Risk

# Disparate Treatment Discrimination

- Perhaps the most obvious form of discrimination.
- Where protected class status is used in the decision-making process.
- In an algorithm, this would manifest as a variable for protected class status being included in the model.
- Disparate Treatment Discrimination:

$$y = f(X, p)$$

Where X are the valid features of the model and p is an indicator of protected class status.

# Proxy Discrimination

There are two types of proxy variables

“Stand-alone proxies”

- A stand-alone proxy is a variable or set of variables so closely related to protected class status that they effectively represent the inclusion of protected class status as a variable in a model (i.e., nearly disparate treatment).

“Predictive Proxies”

- A predictive proxy occurs when features only have predictive power because of their relationship to a protected class.

# Stand-Alone Proxies

Typically measured with some kind of correlation or information measure.

- Information Value: Perform a Weight of Evidence (WoE) transformation (this can be done with PiML) and then calculate Information Value (IV). If the IV is greater than 0.3, this is evidence that the variable has proxy risk.

- Calculate correlations and compare to a threshold – typically 0.80 to 0.95. If the correlation is greater than the threshold, then this is indicative that a variable has proxy risk.

- Another technique that straddles the stand-alone and predictive proxy definitions is measured by calculating the following ratio: Ratio = CORR(x, p) / CORR(x, y). If Ratio > 1 and CORR(x, p) > 0.3, then there is evidence that the variable is a proxy risk.

# Predictive Proxies

In a logistic or OLS model, a predictive proxy can be identified by running the model three times:

1. Run the model across all observations; test coefficients for statistical significance.
2. Run the model for just the protected class; test coefficients for statistical significance.
3. Run the model for just the reference group; test coefficients for statistical significance.

If the coefficient is a statistically significant predictor for all observations, but not for either by-group model, then the variable is likely gaining its predictive power through its ability to differentiate protected class. It is then considered a high risk of being a proxy variable.

Open question: what is the best way to do this with machine learning models, where statistical significance does not get calculated?

# Disparate Impact

- Disparate impact occurs when a valid or “facially neutral” factor causes outcomes to be worse for one group relative to another.

- Evidence of disparate impact does not necessarily mean the model is illegally discriminatory.

- A classic example is a weight-lifting test for firefighters.

- In consumer credit, we often see disparate impact in default model outcomes. Certain minority groups are predicted as being more likely to default on average than other races or ethnicities.

- Disparate impact occurs when:

$$\mathrm{E}(\hat{y}|p=1) > E(\hat{y}|p=0)$$

- Here, the average outcome for the protected group (p=1) is higher than for the reference group (p=0). A higher outcome is assumed to be less favorable (e.g., probability of default).

# Impact, Validity, and Bias

Understanding the source of discrimination in the model

# Handling Disparate Impact

- If a disparate impact is present, this does not necessarily mean that there is illegal discrimination.

- In fact, it is, unfortunately, likely to occur.

- Finding evidence of it does require additional steps to ensure people are being treated as fairly as possible.

- Specifically, one should follow the approach of the "burden-shifting test":

1. Test for evidence of disparate impact.
2. Ensure that the model is valid and that the factors that drive disparities are reasonable predictors.
3. Search for Less Discriminatory Alternative (LDA) models.

- Doing a good job performing the burden-shifting analysis in-house means that regulators and plaintiffs will view you as a less attractive target.

# Measuring Disparate Impact

- Disparate impact is measured as a test of whether unconditioned outcomes differ across groups.

- Disparate impact measures do not incorporate the true outcome. In other words, measures such as relative false positive rates are not measures of disparate impact.

- Commonly used metrics include the Adverse Impact Ratio and Standardized Mean Difference.

$$Adverse\ Impact\ Ratio\ (AIR) = \frac{Selected_{protected\ group}}{Selected_{reference\ group}}$$

$$Standardized\ Mean\ Difference\ (SMD) = 100 * \left(\frac{\hat{Y}_{protected\ group} - \hat{Y}_{reference\ group}}{\sigma_{\hat{Y}}}\right)$$

# Model Validity

- Model validity concerns itself with the issues of whether a model is using the right data to predict the right things for a given usage.

- This is important for assessing model fairness because a model that is not valid is unlikely to be fair to members of certain groups.

Important types of validity

- Face
- Construct
- Criterion
- Content

# Differential Performance and Prediction

Differential Performance is found when a model does a better job (that is, it is more accurate) in measuring the outcome for certain groups relative to others.

- Differential performance can be measured through metrics such as relative AUCs.
- There are non-discriminatory reasons why there may be evidence of differential performance: differences in AUCs are not dispositive indicators of discrimination.

Differential prediction or bias occurs when the relationship between the model's predictions and the true outcome being measured is not consistent across groups, leading to under- or over-prediction for certain groups.

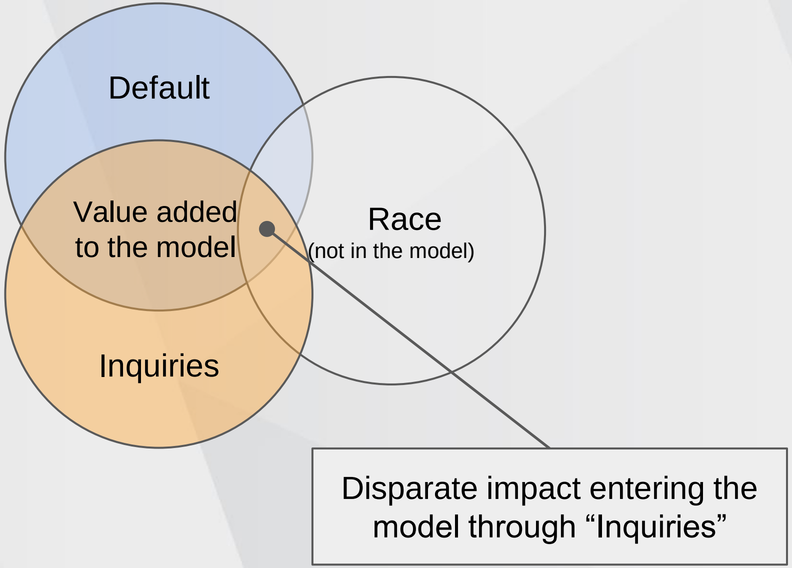
- Differential prediction can be measured by looking at relative model residuals.

# Mitigating Discrimination

Searching for Less Discriminatory Alternative (LDA) Models

# Making Fairer Models via Feature Selection

## Model 1 - Includes Inquiries



Default

Value added to the model

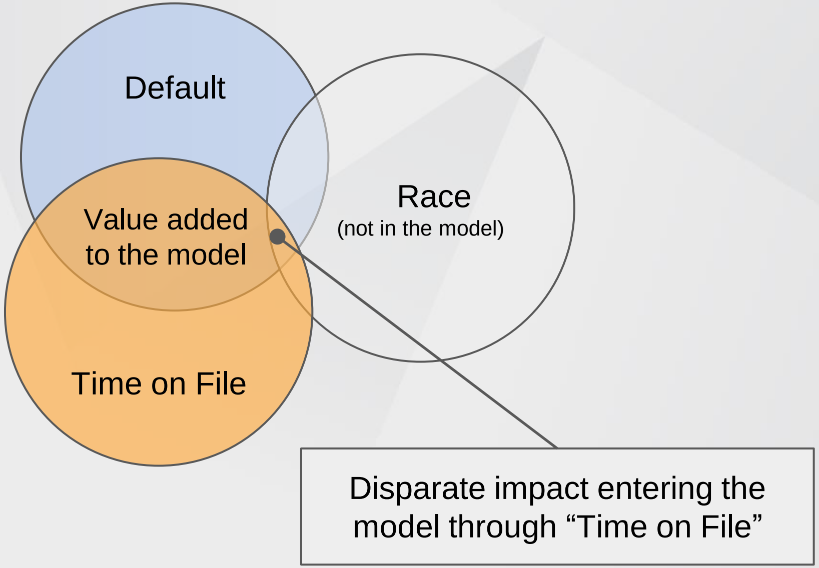
Inquiries

Race

(not in the model)

Disparate impact entering the model through “Inquiries”

## Model 2 - Includes Time on File



Default

Value added to the model

Time on File

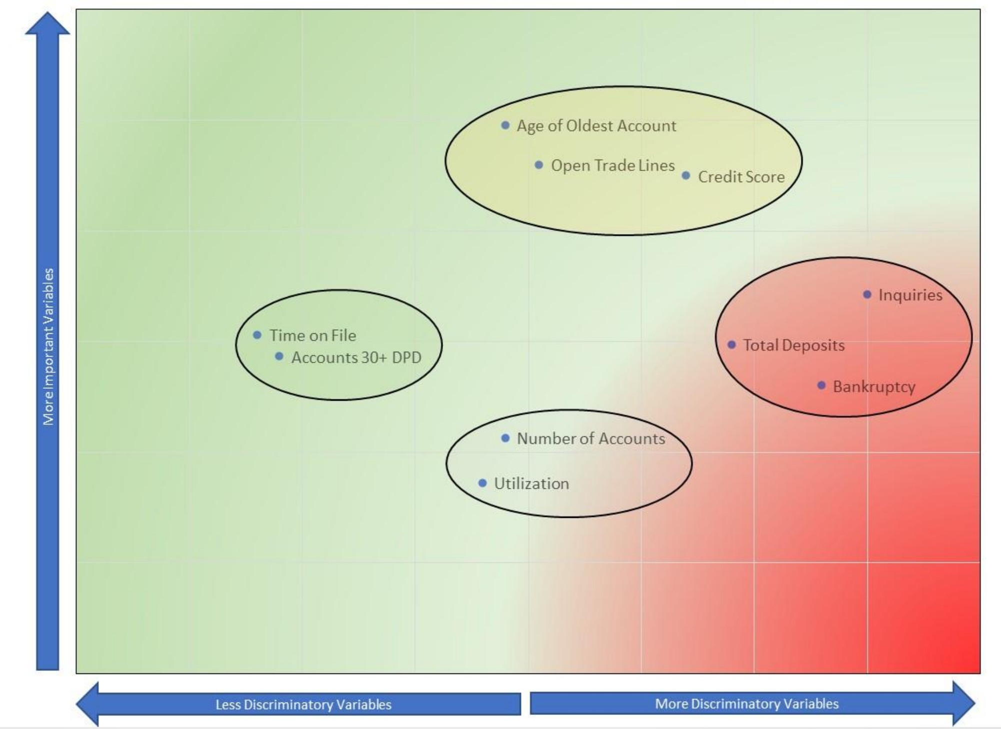
Race

(not in the model)

Disparate impact entering the model through “Time on File”

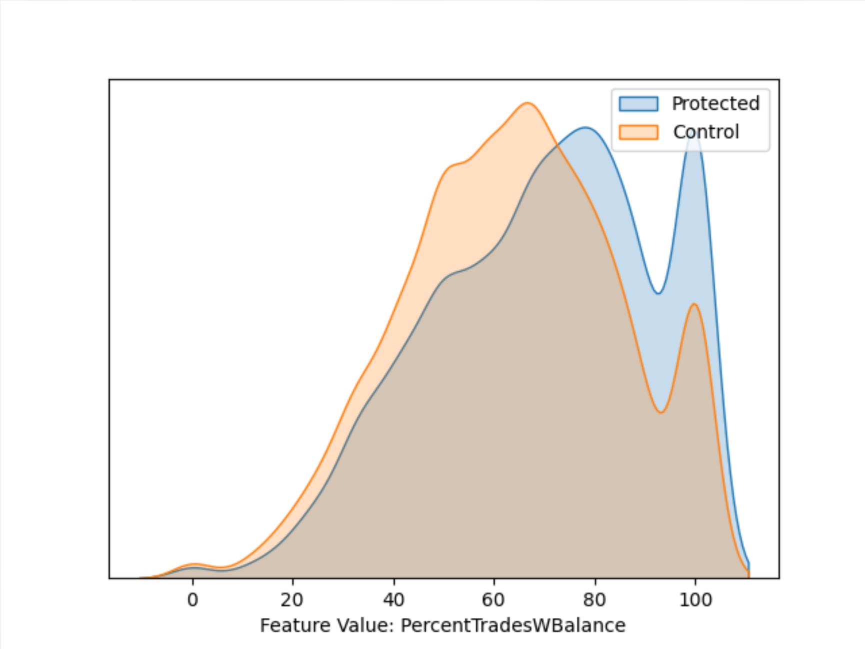
# Debiasing Models Using Explainable AI

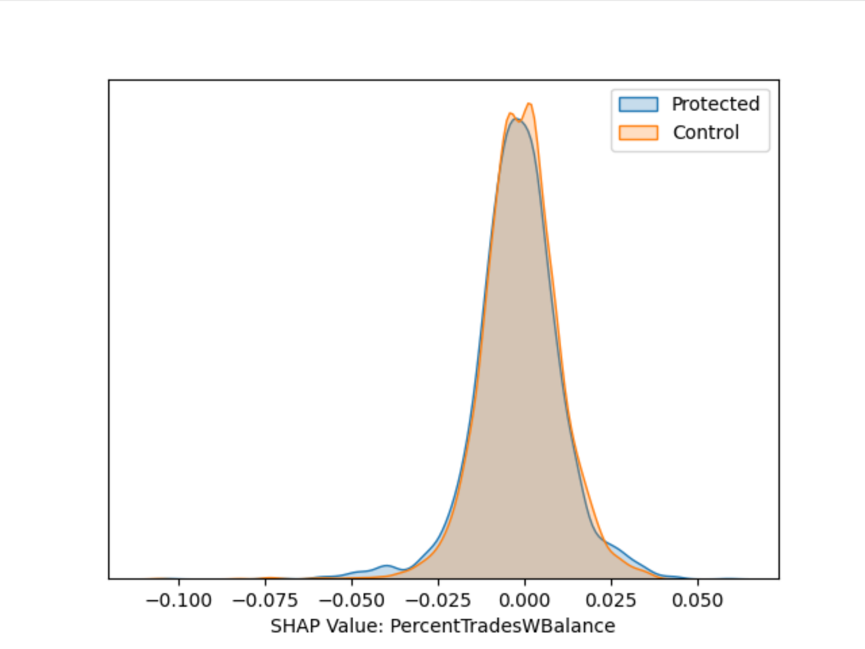
- Identify the variables that are causing discrimination
- Identify the variables that are highly important
- Re-create the model, focusing on important, but less discriminatory variables



# Debiasing Models Using Explainable AI

Percentage of Trades with a Balance





# Less Discriminatory Alternative Search

The third prong of the burden-shifting test requires that a valid model that shows evidence of disparate impact undergo testing to search for less discriminatory alternative models.

- The search needs to be reasonable but does not have to be onerous.

- Traditionally, it was done by swapping in and out a few variables.

- Other more effective techniques are supplanting this method.

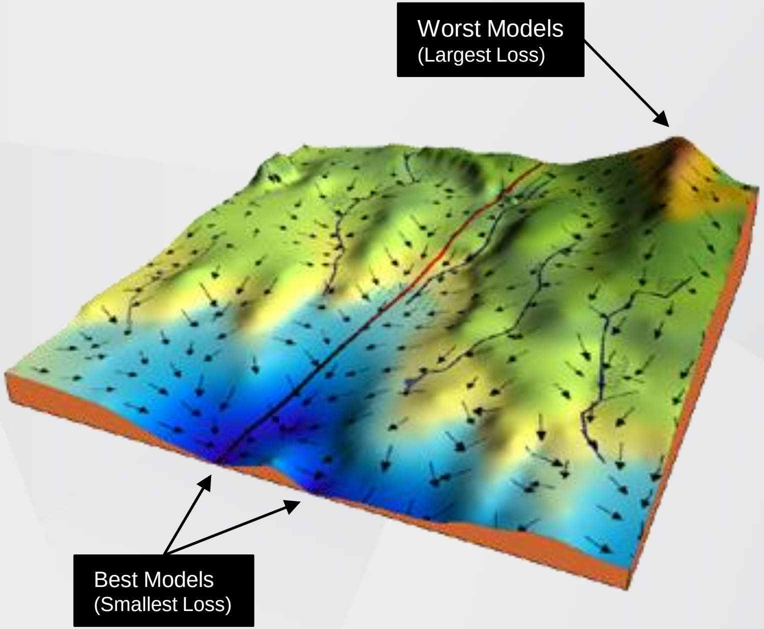
Options include:

- More intelligent feature selection
- Hyperparameter tuning
- Incorporating protected class characteristics directly into model training

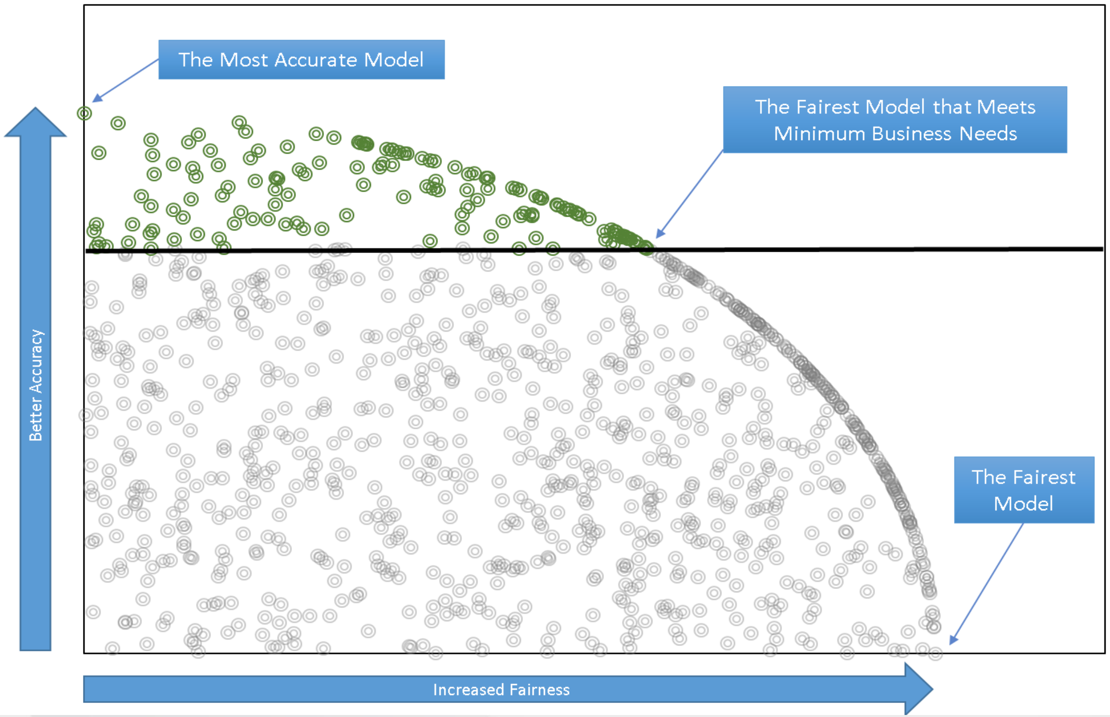
# The Multiplicity of Good Models

- Because machine learning models have such flexibility, **more than one model may meet all necessary model governance requirements**

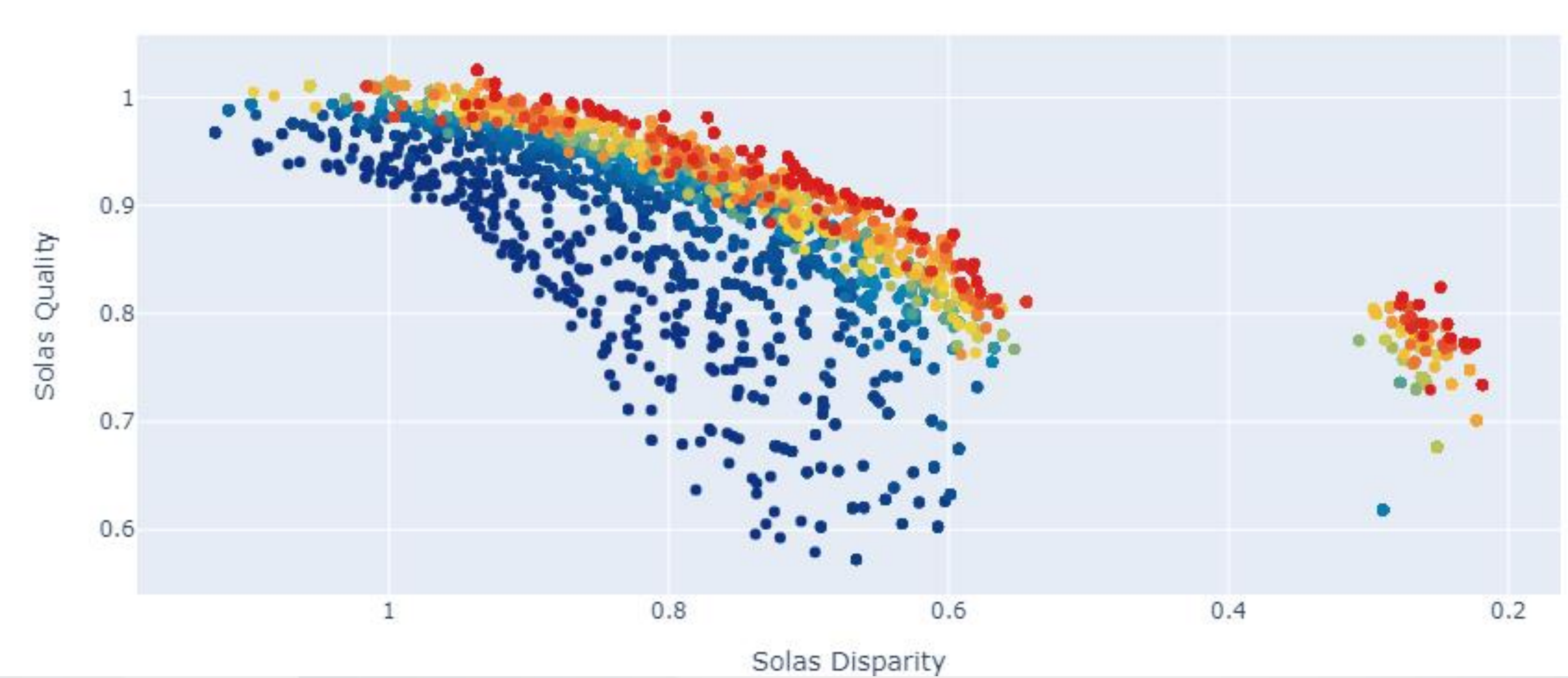
- This gives us the opportunity to optimize on more than one metric: **fairness**



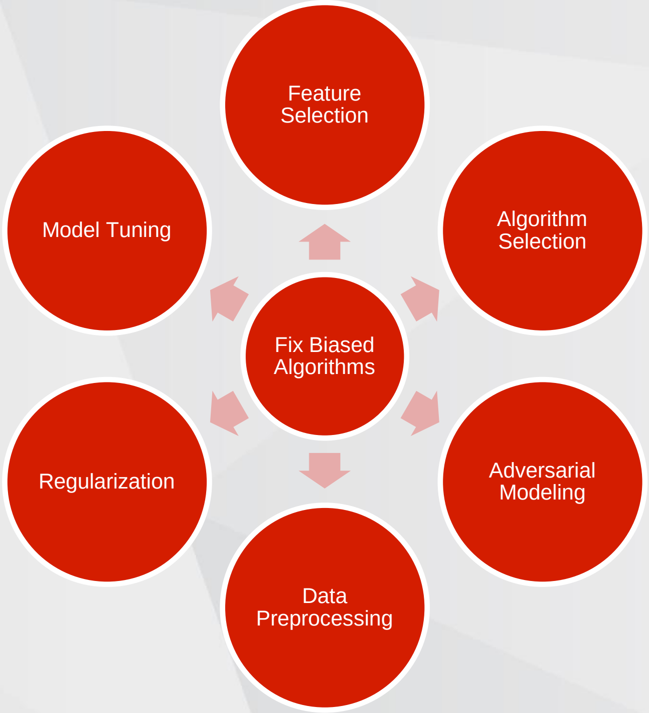
# Using AI to Fix AI: the Pareto Frontier



# Using AI to Fix AI: A Real-world Example



# Available techniques for mitigating discrimination



- Many options
- Choice matters!
  - Good intentions may lead to harm
- Most open-source implementations have limited in-production capability
  - IBM Fairness 360
  - Fairlearn
  - Aequitas
- Open-access testing capabilities
  - SolasAI (https://github.com/SolasAI/solas-ai-disparity)

# Pitfalls in Fairness Analyses

- **Putting garbage or risky data in a model**
  - Are the features discriminatory?
- **Not measuring disparities considering actual outcomes**
  - Ensure thresholds correspond to practice
- **Getting compliance advice from people with no compliance or business experience**
  - Expertise is essential
- **Running the cool new de-biasing algorithm**
  - But does it follow regulatory compliance?
  - What is it accomplishing?
- **Getting too attached to the model you’ve chosen**
  - There are almost certainly many other similar models

# Thank you!

Nicholas Schmidt
nick@solas.ai