FinRegLab

Machine Learning Explainability & Fairness
Insights from Consumer Lending

FinRegLab in collaboration with Professors Laura Blattner and Jann Spiess

August 31, 2022

Research Purpose

FinRegLab collaborated with researchers from the Stanford Graduate School of Business to conduct empirical research on **the capabilities, limitations, and performance of available proprietary and open-source model diagnostic tools.** We aim to:

- Provide **independent evidence** about the capabilities, limitations, and performance of tools designed to help lenders understand and manage machine learning underwriting models

- Develop and implement a research methodology that:

  - Focuses on **model diagnostic tools currently being used** in consumer lending
  - **Approximates lenders' use** of those tools
  - **Reflects input from diverse stakeholders,** including lenders, advocates, and policymakers
  - Highlights **implications of using machine learning models of varying degrees of complexity**

- Propose a **systematic approach** for evaluating whether and in what circumstances information produced to describe model behavior can be used to satisfy various consumer protection and prudential expectations

- **Inform the evolution of policy, market practice, and technology**

What We Did

FinRegLab, a non-profit research organization, teamed up with Stanford Professors Laura Blattner and Jann Spiess to **evaluate model diagnostic tools from seven technology companies and select open-source tools** in the context of consumer lending.

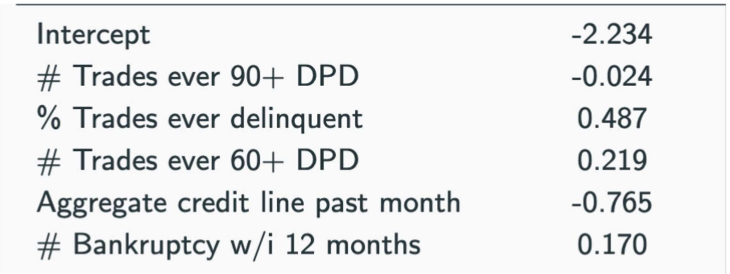
The following companies provided model diagnostic tools for use in this evaluation:

Arthur fiddler H_2O.ai relationalAI SOLAS AI Stratyfy Where AI meets IQ ZEST AI

Motivation

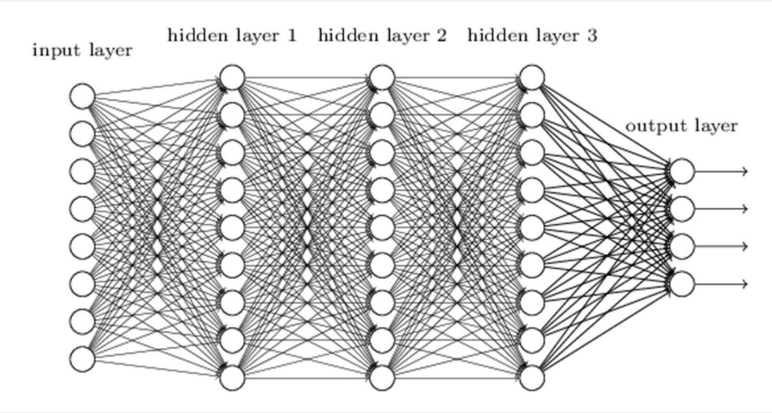
Adoption of machine learning models to extend consumer credit intertwined *with key threshold question*:

How do we solve transparency challenges associated with machine learning models?



Intercept	-2.234
# Trades ever 90+ DPD	-0.024
% Trades ever delinquent	0.487
# Trades ever 60+ DPD	0.219
Aggregate credit line past month	-0.765
# Bankruptcy w/i 12 months	0.170

Logistic Regression Model

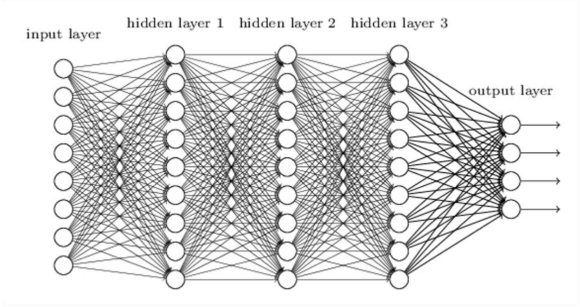


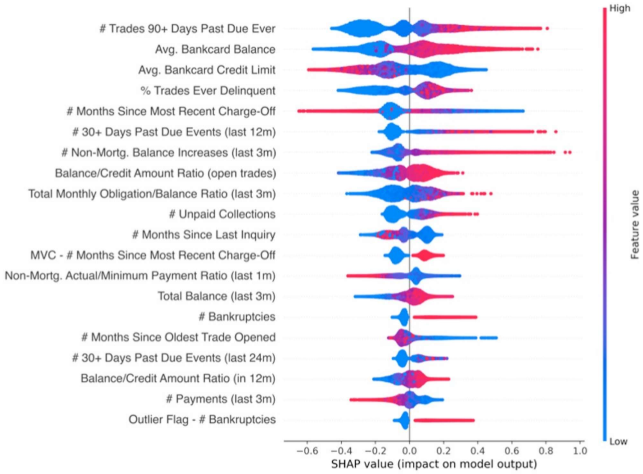
Neural Network

Transparency Challenges in Consumer Credit

1. Why was **this loan applicant** rejected based on the model?
2. Why does the model have different behavior for **different protected class groups**?
3. What drives the **deterioration in model performance** in an out-of-time context?

Our focus is on how well **model diagnostic tools applied "*post hoc*"** help lenders answer these questions.

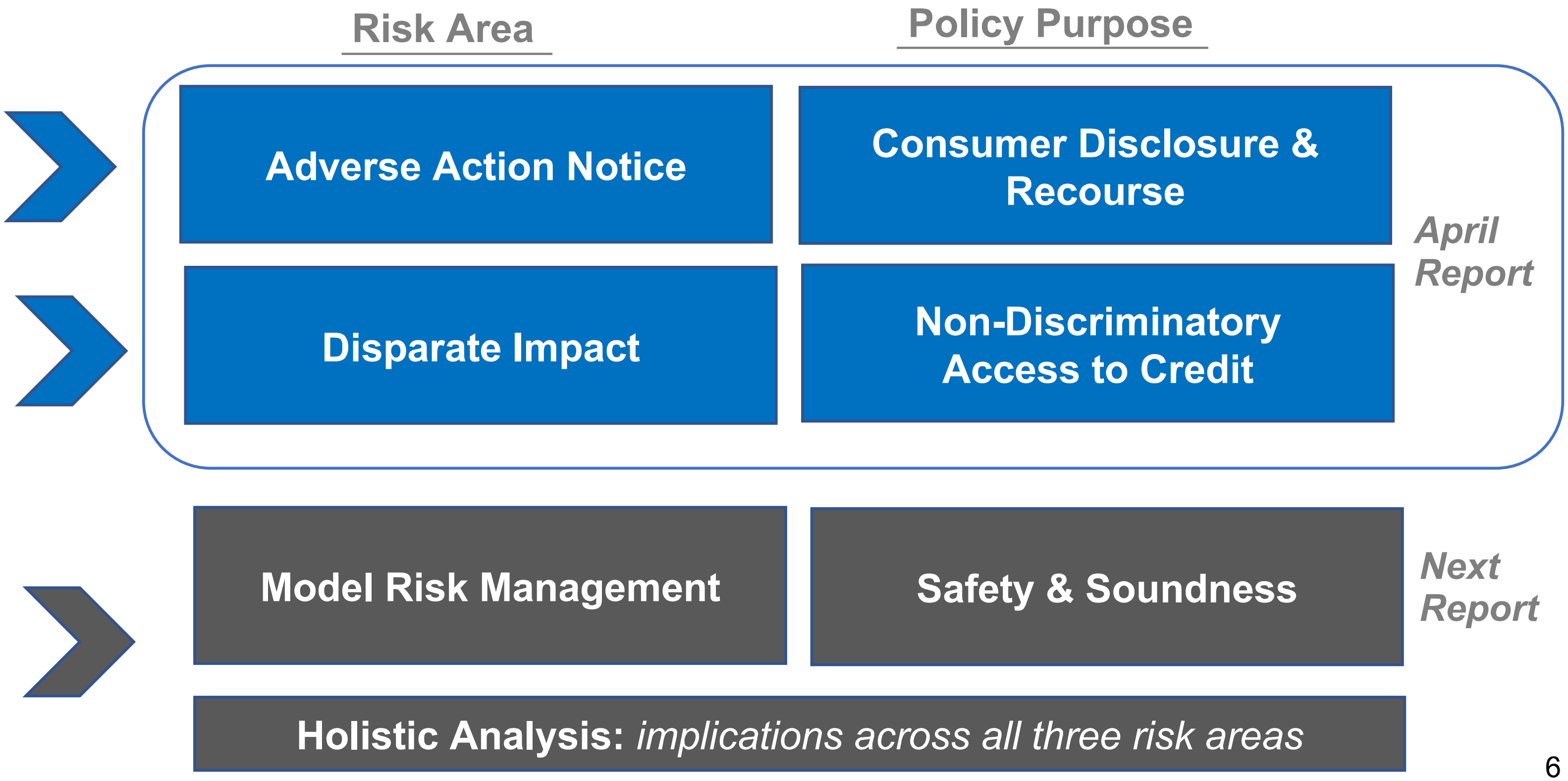




FinRegLab

How We Designed the Research

Transparency is a ***means to an end***



Risk Area	Policy Purpose	
Adverse Action Notice	Consumer Disclosure & Recourse	*April Report*
Disparate Impact	Non-Discriminatory Access to Credit	
Model Risk Management	Safety & Soundness	*Next Report*
Holistic Analysis: *implications across all three risk areas*		

FinRegLab

Key Findings

We found that certain model diagnostic tools can generate important information about the model's behavior in context of specific consumer protection requirements regarding disparate impact and adverse action notices. These tools work when firms:

- **Pick the right tool for the task at hand**
- **Interpret explanatory information from the tools in light of correlations**

Nevertheless, **explainability alone does not deliver less discriminatory alternatives for ML models** (or feasible paths to acceptance) – *instead, a broader range of tools is needed to achieve that.*

Use of secondary tools **adds another set of consequential decisions** to the many choices that firms make when developing and using machine learning underwriting models that can be used in compliance with existing requirements and expectations.

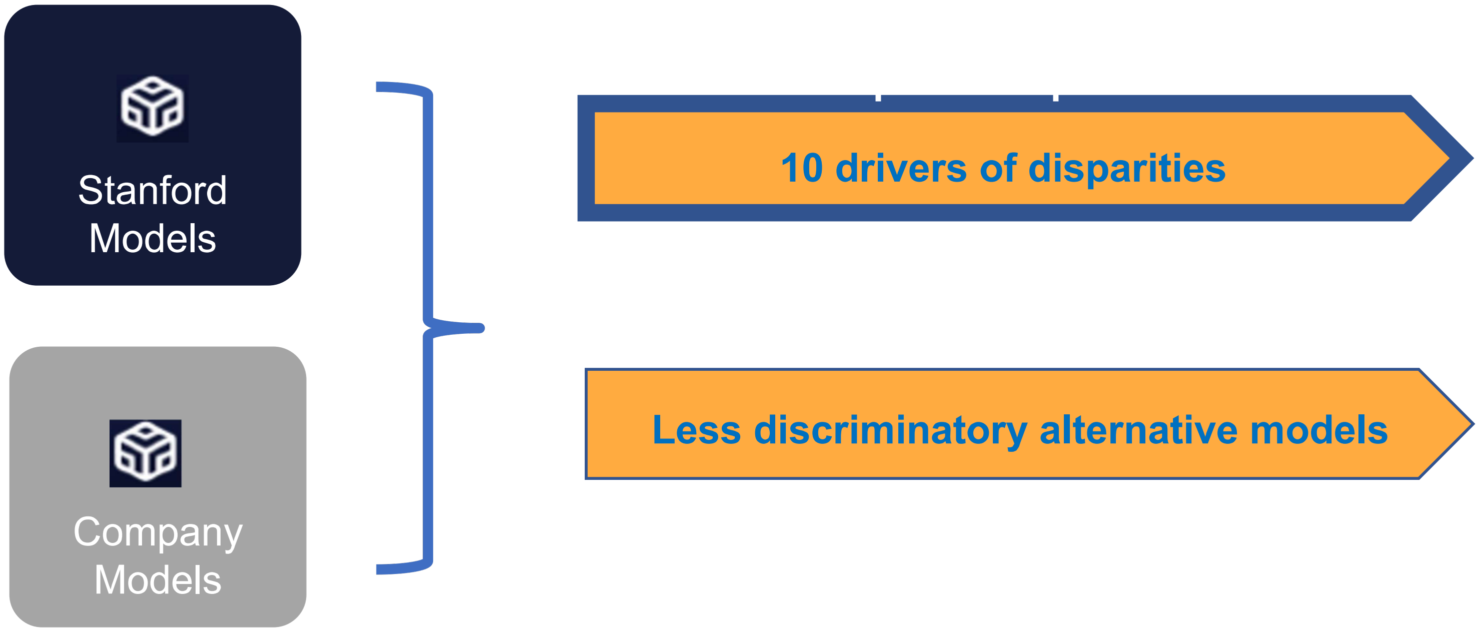
- **The quality of the chosen tool and judgments about its use mattered more than model complexity** in determining whether information produced to describe model behavior was usable for compliance purposes.

FinRegLab

Key Findings: *Disparate Impact*

What Participating Companies Did

Each participating company used its tools to accomplish the following tasks and in so doing, used different approaches to generating information about model behavior.



Stanford Models

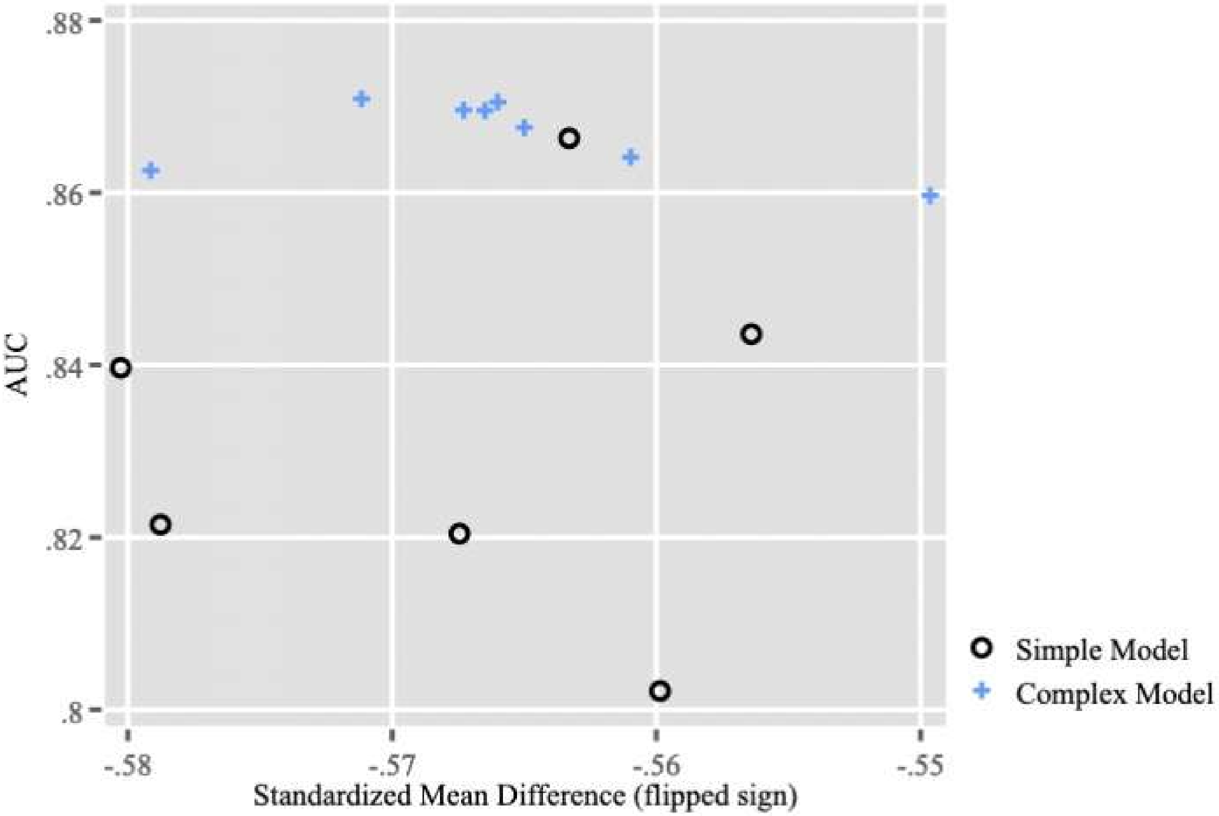
Company Models

10 drivers of disparities

Less discriminatory alternative models

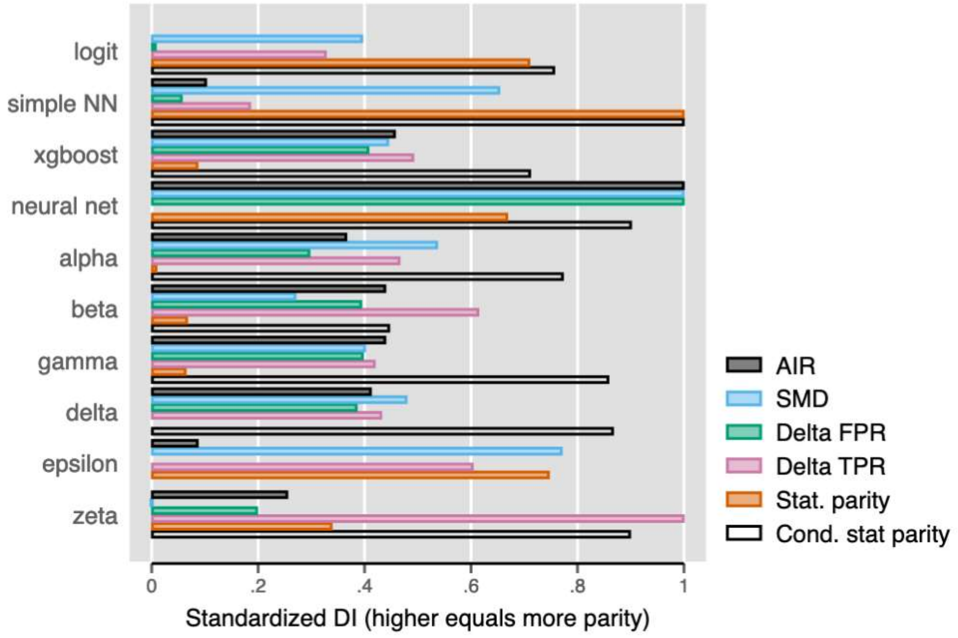
Fairness and Performance Properties on Test Data

Complex models (both Stanford and company built) outperformed on predictive performance AND fairness metrics

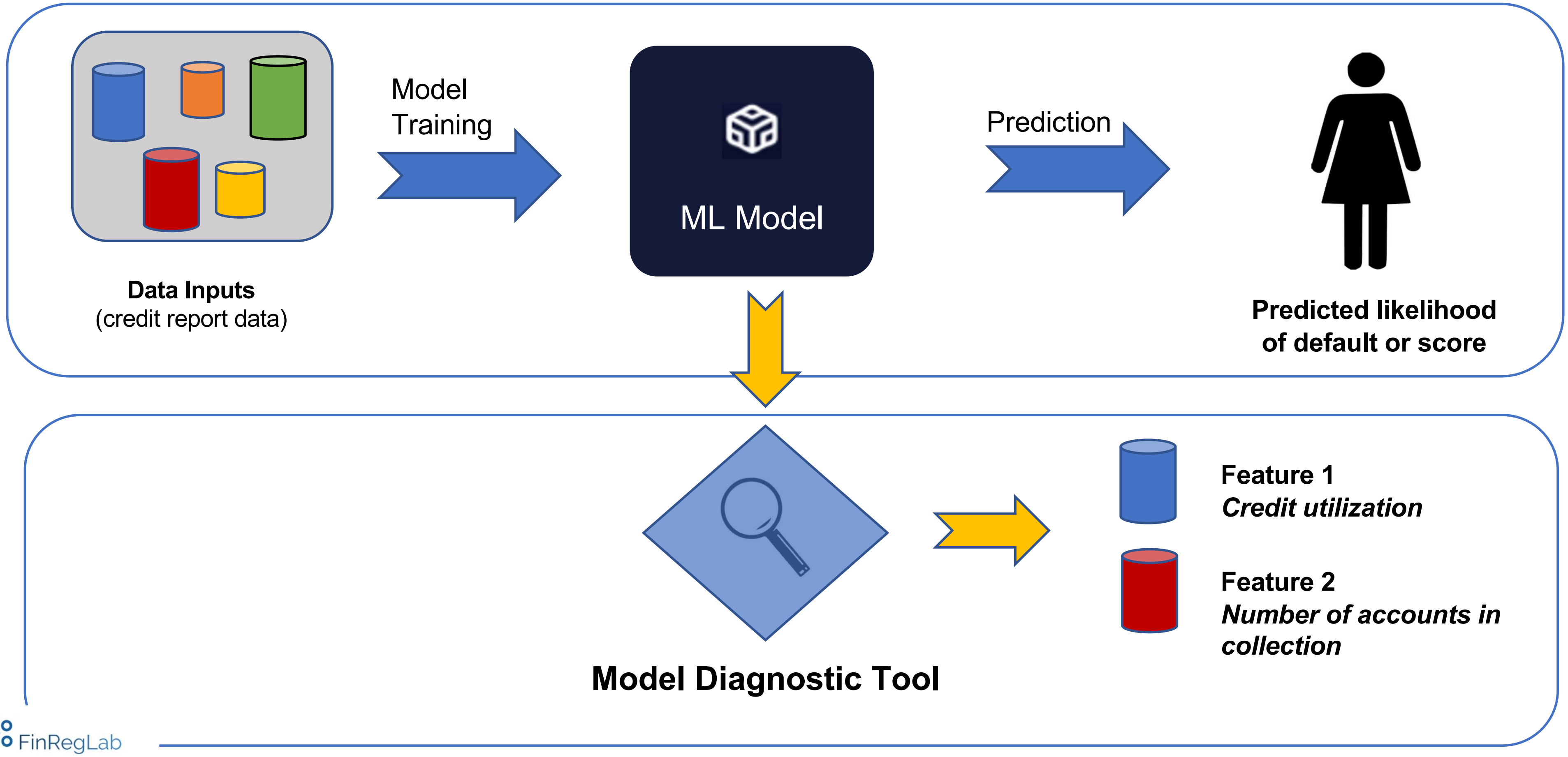


Fairness Properties

No single model performs best on all fairness metrics (reflecting well-known impossibility results)



Experiment Layout



Data Inputs
(credit report data)

Model Training

ML Model

Prediction

Predicted likelihood of default or score

Model Diagnostic Tool

Feature 1
Credit utilization

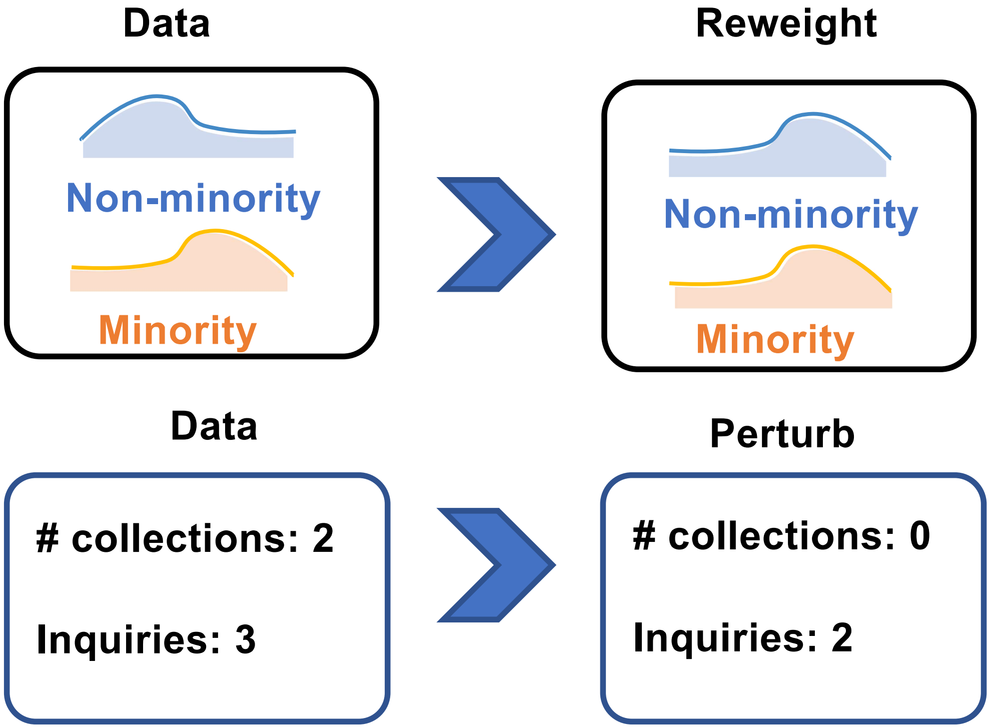
Feature 2
Number of accounts in collection

Fidelity: Drivers of Disparities

Companies identified 10 drivers of disparities

Reweighting Test

Perturbation Test



Data

Reweight

Non-minority

Minority

Data

Perturb

collections: 2

Inquiries: 3

collections: 0

Inquiries: 2

Fidelity: Results

- Large differences between tools
- Note: Reweighting implicitly considers correlated features.
- Similar performance for simple and complex models

REWEIGHTING TEST

	Change in AIR (larger is better)	Beat random (100% is best)
Best tools	33pp	100%
Worst tools	9pp	0%

PERTURBATION TEST

	Beat benchmark (100% = best)	Beat random (100% = best)	Beat correlated (100% = best)
Best tools	100%	100%	100%
Worst tools	0%	0%	0%

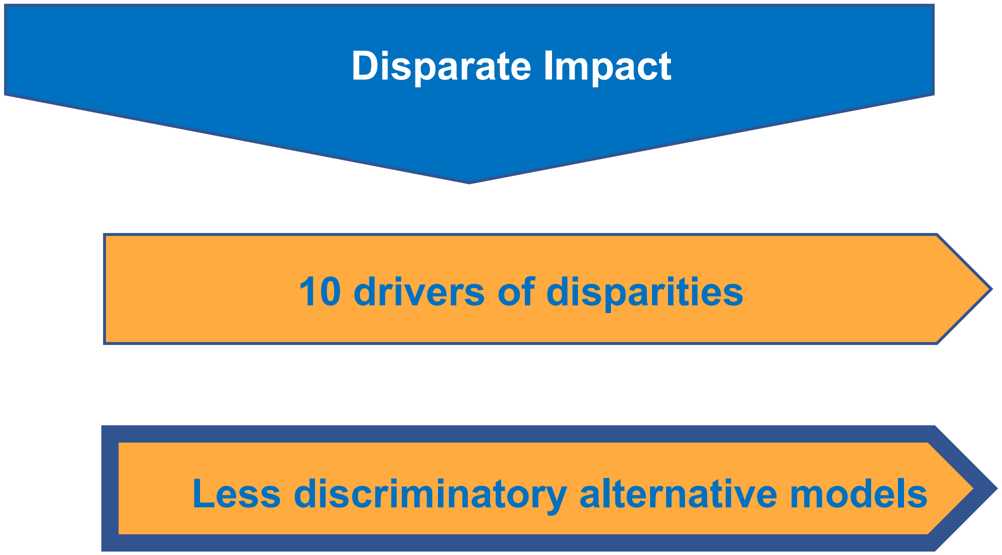
Drivers of Disparity

- Some, but not all, tools have **high fidelity** for both logit and complex models.

- **High fidelity tools produce more consistent information for both logit and complex models** than low fidelity tools, in particular when feature-level explanations are aggregated.

What Participating Companies Did

Each participating company used its tools to accomplish the following tasks and in so doing, used different approaches to generating information about model behavior.

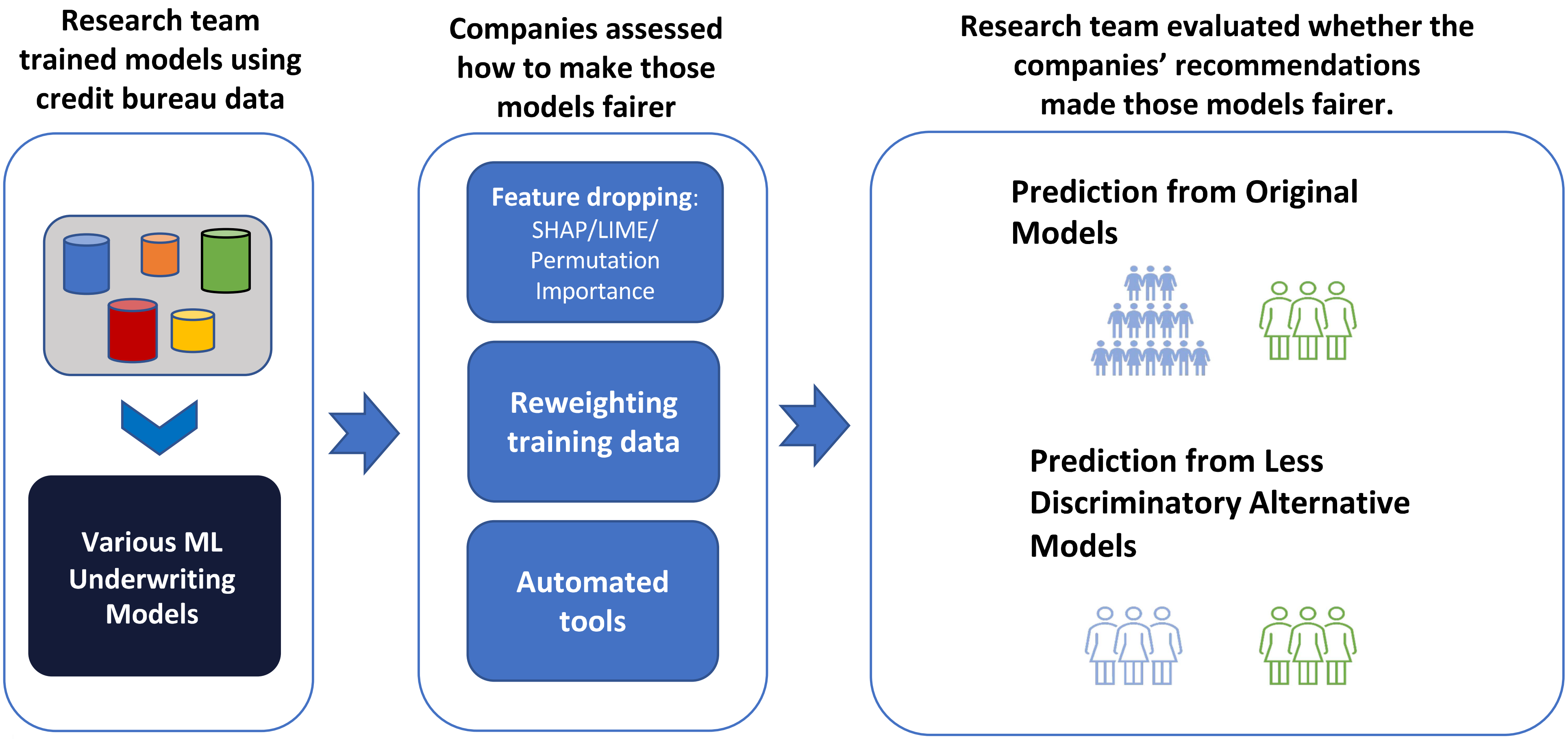


Disparate Impact

10 drivers of disparities

Less discriminatory alternative models

Experiment Layout: LDA Search



Research team trained models using credit bureau data

Various ML Underwriting Models

Companies assessed how to make those models fairer

Feature dropping: SHAP/LIME/ Permutation Importance

Reweighting training data

Automated tools

Research team evaluated whether the companies' recommendations made those models fairer.

Prediction from Original Models

Prediction from Less Discriminatory Alternative Models

Search for Less Discriminatory Alternatives (1/2)

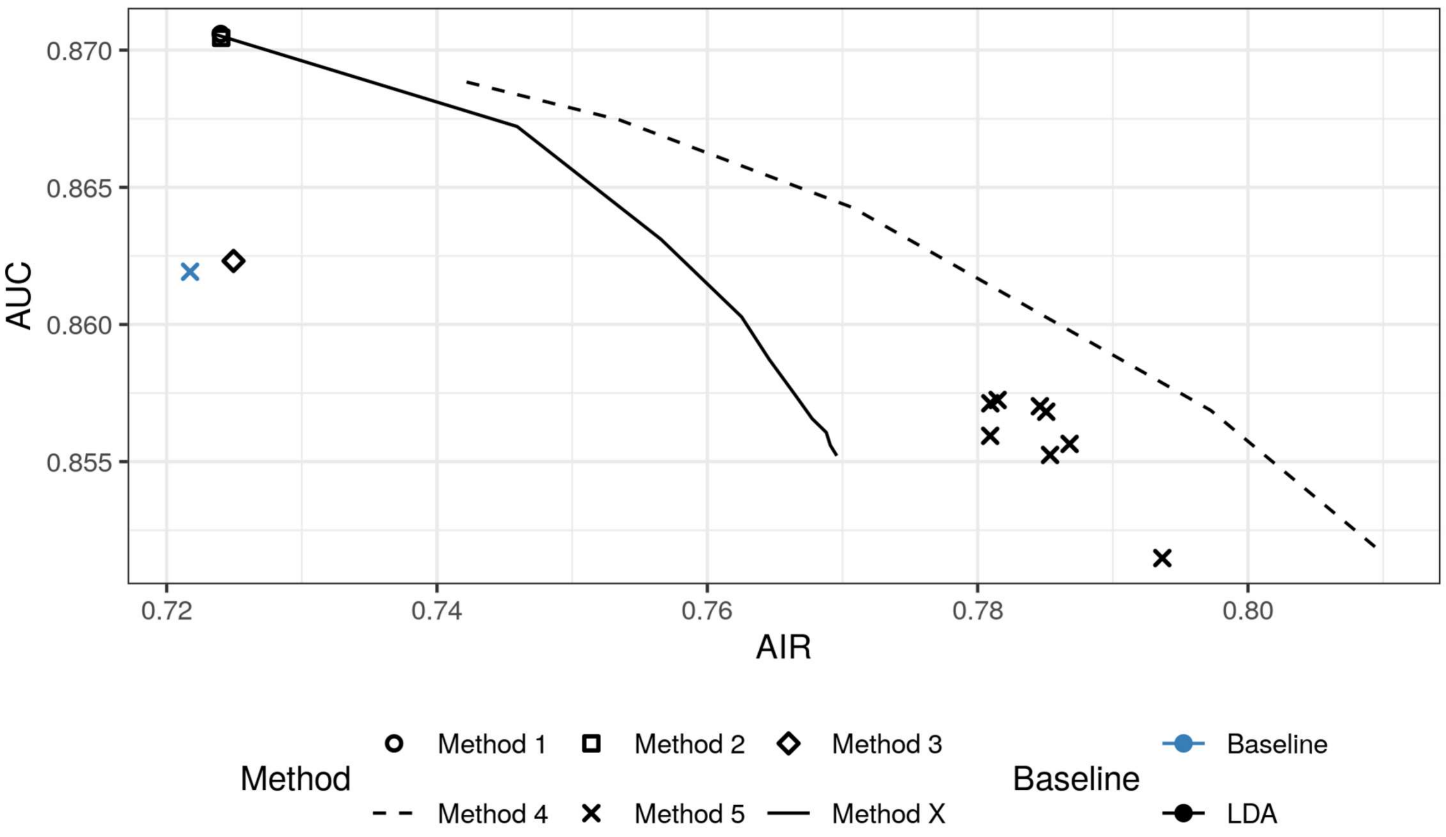
Baseline Metrics

Used to assess improvements in recommended less discriminatory alternatives when measured against XGBoost models developed by the research team.

- Area Under the Curve (AUC)
  - 0.87
- Adverse Impact Ratio (AIR)
  - 0.73

Search for Less Discriminatory Alternatives (2/2)

XGBoost results

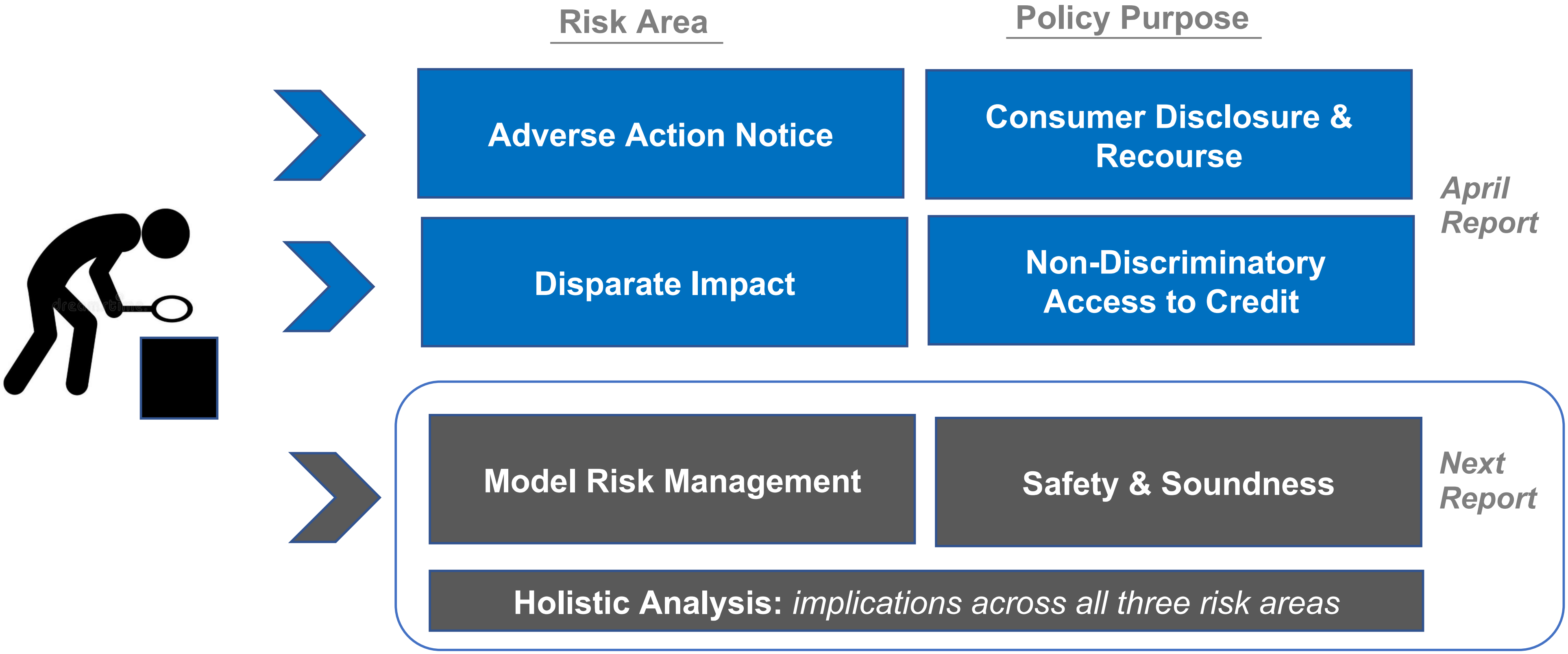


FinRegLab

Usability + LDA Search: Key Points

- **Automated approaches outperform** feature-drop and reweighting strategies.
- Our results generalize **generalized well** to:
    - Different populations
    - Different model types
- **Results varied depending on model type and fairness metric used**; no single automated approach is always best.
- **More fairness is possible** with ML approaches, but may come at **some performance cost**.

Next Steps



Risk Area

Policy Purpose

Adverse Action Notice

Consumer Disclosure & Recourse

Disparate Impact

Non-Discriminatory Access to Credit

April Report

Model Risk Management

Safety & Soundness

Next Report

Holistic Analysis: *implications across all three risk areas*

FinRegLab