Algorithmic Discrimination and Input Accountability under the Civil Rights Acts

Robert Bartlett (UC Berkeley Law),
Adair Morse, Richard Stanton & Nancy Wallace (UC Berkeley Finance)

+ a finance application of the ideas not yet a manuscript

Adair Morse

August, 2020

Contribution

I. SETTING

Lenders use 1,000s of variables for algorithmic profiling.

Challenge:

How to implement Civil Rights Act for determining what is legal statistical discrimination

II. OUR CONTRIBUTION

Following the Civil Rights Act, an implementation framework emerged from **Supreme Court caselaw + legislation (Congress)** that provides explicit legal framework.

Our Contribution:

- **combining legal framework &**
- **economic fundamental model**

We put these pieces together to lay out what it means for lending algorithms to be accountable under discrimination law.

How **Economists** Think about Discrimination

TASTE-BASED DISCRIMINATION

- A decision-maker derives utility from discriminating against a protected group (Becker `57)

- Should not persist in the long run because of competition

  Taste-Based Discrimination is costly

- De facto: persists

STATISTICAL DISCRIMINATION

- A decision-maker does not observe a business necessity variable (e.g.,cash flow variables of credit risk).

- ***Direct:*** Uses an average for a group of people (Arrow, 1973, Phelps, 1972) based on protected category

  Statisitcal Discrimination profit maximizes

- De facto use of statistical discrimination: mostly ***indirect stat discrimination***:

  = using averages over a non-protected variable (not "black" but "goes to Ivy League college")

How do Lenders think about discrimination?

Lender : a lender with 1,000s of variables wants to use machine learning (ML) to do credit scoring without discrimination

Corp. Lawyers: *“To avoid discrimination, apply a 'least discriminatory’ approach”*

How?

1. Define the business necessity for using proxy variables
   - Courts: in lending = “credit risk” (not expected profit of loan)
2. Run predictive accuracy models of default
   - Default is (an imperfect) ex post measure of ex ante credit risk
3. Then, (especially if resulting outcomes are disparately applied against a protected category), show that the algorithm uses the least discriminatory predictive model for a given level of predictive accuracy

How the Law Thinks about Discrimination

The mapping of the law to economists' thinking is clear on the below:

1. Make taste-based discrimination illegal

2. Make sure technology does not implement the direct form of Arrow/Phelps discrimination
   - i.e.: allowing lenders to score by a protected category or a "highly correlated" variable
     - Protected category: race, ethnicity, gender, etc.
     - Highly-correlated = hair styles, redlining, etc.

How the Law Thinks about Discrimination

But the law is not quite so simple as 1 and 2:

1. Make taste-based discrimination illegal
2. Make sure technology does not implement the direct form of Arrow/Phelps discrimination

Disparate treatment

What about indirect statistical discrimination??

This is where our contribution comes in….

Disparate Impact?

U.S. Title VII of the Civil Rights Act of 1964

An unlawful practice for an employer

1. "to … discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, sex, or national origin; or
2. to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities … because of such individual's race, color, religion, sex, or national origin."

A long-standing challenge: How do you implement this in a setting where discrimination may be unintentional?

Burden- Shifting Framework

Caselaw that was later codified as implementation law

Original frame from Supreme Court:

- *Griggs v. Duke Power Co*

Codified by Congress:

- *Civil Rights Act of 1991*

Important Caselaw from Supreme Court*:*

- *Ricci v. DeStefano*
- *Dothard v. Rawlinson*

- Original application is in context of employment decisions.
- Credit and housing decisions adopted this interpretation of discrimination and this framework explicitly in Equal Credit Opportunity Act and Fair Housing Act

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes "observed statistical disparities" across members of protected and unprotected groups.

- If plaintiff successful…

Second Burden: The defendant must then "demonstrate that the challenged practice is *job related for the position in question* and consistent with business necessity."

- If defendant successful…

Third Burden: Plaintiff must show that an equally valid and less discriminatory practice was available that the employer refused to use

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes "observed statistical disparities" across members of protected and unprotected groups.

Second Burden: The defendant must then "demonstrate that the challenged practice is *job related for the position in question* and consistent with business necessity."

Third Burden: Plaintiff must show that an equally valid and less discriminatory practice was available that the employer refused to use

#1: This is where the least discriminatory approach comes from

#2: But it does not excuse the defendant from satisfying Second Burden

Dothard v. Rawlinson

A California Prison wanted to hire prison guards

- Determined that a job-required necessity is strength (legitimate)
- Could not measure strength of applications, so used proxy of height
- A group of female applicants sued and won

Court:

- Indeed strength is legitimate as business necessity and height predicts performance
- But the strength needed is a specific strength and the height measurement penalizes females beyond the business necessity

Lending version

Business necessity is creditworthiness (i.e.: ex ante credit risk)

What is the economic fundamental model describing this business necessity to identify targets?

Imagine writing down a structural model of expected cash flow available for repayment.

- Target variables : Life-cycle or permanent income variables…
  - Income, income growth, wealth, cost of capital, cost of consumption, existing debt, etc

Employment version

Business necessity is skills required for the job.

What is the economic model?

In most applications, the model is simplified to linear function

- Skills required = f(strength, dependability, cognitive and psychological IQ)

Comback to this slide
Dothard v. Rawlinson

A California Prison wanted to hire prison guards

- Determined that a job-required necessity is strength (legitimate)
- Could not measure stre
- A group of female appl

How do we turn this into a systematic way for statistical testing?

Court:

- Indeed strength is legitimate as business necessity and height predicts performance
- But the strength needed is a specific strength and the height measurement penalizes females beyond the business necessity

Dothard v. Rawlinson: IAT

- Econometrician Version

1. Decompose height into that which predicts the target strength and a residual
2. Test if the residual is still correlated with female:

$$Height_i = \alpha \cdot Strength_i + \varepsilon_i$$

Test: $\varepsilon_i \perp gender.....$ $regress\text{:}\ \ \varepsilon_i = \beta_0 + \beta_1 gender$

Proxy height fails ⇔ $\beta_1 \neq 0$

If so, exclude height as only legitimate business necessity

We call this the *Input Accountability Test*

Dothard v. Rawlinson: IAT

If doing a 1,000 variable estimation, the proxy input variables may not aim at a single target, but rather the overall business necessity – credit risk

1. Decompose the input variable into that which predicts any of the fundamental model targets
2. Test if the residual is still correlated with female

$$Ivy\ League_i = \alpha_1 \cdot Income_i + \alpha_2 \cdot CreditScore_i + \alpha_3 \cdot Wealth_i$$
$$+\alpha_4 \cdot EIncomeGrowth_i + \cdots . + \varepsilon_i$$

Test: $\varepsilon_i \perp race.....$ $\quad regress\text{:}\ \varepsilon_i = \beta_0 + \beta_1 race$

Proxy height fails ⇔ $\beta_1 \neq 0$

Challenges of the IAT

1. Unobservability of Target
   - Kleinberg, Ludwig, Mullainathan, Sunstein (2019): *training datasets*

2. Measurement Error in Target

$$Strength_i^* = Strength_i + \mu_i$$

$$Height = \alpha \cdot Strength_i^* + \zeta_i$$

$$\zeta_i = -\mu_i + \varepsilon_i$$

3. Standard errors as n grows large.

A fix instead of exclude?

Question: Why can’t we just fix the scoring by a protect group to de-bias?

- Pope and Sydnor (2011)

Answer: It only works on average, not for individuals. The law is about individuals

Answer: It is illegal. *Ricci v. DeStefano:*

New Haven wanted to discard the results of an “objective examination” that sought to identify city firefighters who were the most qualified for promotion because there was statistical racial disparity in the results against a minority group. A group of white and Hispanic firefighters sued, alleging that the city’s discarding of the test results constituted race-based disparate-treatment.

Court ruled for plaintiff… no discarding

Why: Can’t use protected class variables in a decision => could cause disparities

Setting - revisited

- How do **Economists** think about Discrimination?
- How do Finance practitioners (**Lenders**) think about Discrimination?
- How does the **Law** think about Discrimination?
- How do **Fairness Arguments** think about Discrimination?
  - Note: this is not a legal version of discrimination. Important to distinguish fairness from discrimination law. (Fairness is also important.)

Fairness: Ventilators

- Hospitals consider **triage algorithms** to allocate based on **LT survival**
- **Sequential Organ Failure Assessment SOFA**: degree dysfunction, 6 organs

Problem: Legacy of structural racism and inequality => Black and Latinx Americans higher rates of hypertension, diabetes, chronic kidney disease, pulmonary disease, etc.

- Under IAT: If LT Survival is business necessity, then the differential rates of whites getting ventilators is justified.
- Fairness Arguments: Would need POLICY (legislation) to re-define the business necessity target to accounts for the structural inequities that contribute to the racial and ethnic disparities in outcomes.

Credit Score biases: Fairness or Discrimination Law?

What if the credit score is biased against people of color because

- They have less chance to build credit histories because of structural inequities
- They were turned down for credit because of discrimination, conditional on observables in credit application
  - Butler, Mayer, Weston on auto loans
  - Bartlett, Morse, Stanton, Wallace (2019) on yes/no in GSE market

- Giles and Spiess
  - "discrimination stress testing" in lending ➔ (my relabel) "Fairness stress testing"

More needs to be done

Implementation in Finance – not yet a manuscript

Motivation

- U.S. household debt: $14 trillion
  - Increase of $1.3 trillion from peak in 2008 (NY Fed)
  - If annual debt turnover is 15%

- New float of recent years ~$2.2 trillion per year
- Of this, how much algorithmically-decided based on1,000s of proxy variables?

- Bartlett, et al (2019): 45% of lenders in mortgages have fully automated lending (in 2018)

Footprints & Discrimination

Question:

How can the use of machine learning in credit profiling avoid being inadvertently discriminatory?

Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2019),

Bartlett, Morse, Stanton, and Wallace (2019)

Data

- Data from a consumer lender in Eastern Europe
- 300,000 consumer loans

Unique:

- 124 variables (many of the them categorical)
- Can be made “long” into 1,000s of variables even without interactions
- Dataset contains default (the target)

Step 1 – Looking for footprints

Footprints of creditworthiness literature (abridged)

- Berg, Burg, Gombovic, and Puri (2019) :"digital footprints" type of device (tablet, computer, phone), operating system (Windows, iOS, Android), and email provider predicted default rates among the customers of a German lender.
- Bjorkegren and Grissen (2019) mobile phone usage data
- Vissing-Jorgensen (2010) : Consumer goods products people buy

Types of Variables

1. Fundamentals (cash flow, wealth, cost of capital)
2. Occupation
3. Goods
4. Shelter
5. Family Life
6. Soft Info Applying
7. Soft Info Credit

ROC Analysis

Logit (Default) = fundamentals + (iteratively, then all)

1. Occupation
2. Goods
3. Shelter
4. Family Life
5. Soft Info Applying
6. Soft Info Credit

Fundamental Variables

	Mean	StDev		Mean	StDev
Income monthly	168,797	237,125	Missing data Credit Bureau	0.1350	0.3417
Credit Amount	599,028	402,494	# Outstanding Loans	4.3184	10.5095
Payment Amount	27,109	14,494	Prior Loans Delinquent %	0.0054	0.0312
payment_to_credit	0.0537	0.0225	How Delinquent, if any	0.0089	0.0851
payment_to_income	0.1809	0.0946	Ontime Prior Payments, if any	0.1371	0.2522
Homeowner	0.6937	0.4610	Percent of Prior Loans Closed, if any	0.0991	0.2089
Credit Score Max	0.6159	0.1561	Remaining Days on Last Issue	-928.0	644.8
Cedit Score Min	0.3996	0.1874	Days Since Last Issue	-419.3	526.3
# Credit Bureau Requests	0.2313	0.8568	Own Car?	0.3401	0.4737
			Age of Car, if any	0.3418	0.7508

Note: Monetary units are disguised.

ROC Analysis … Columns adding Proxies

Do the proxies add to the ROC?
(Guided Lasso Optimizing)

Dependent Variable: Default
Model: Logit

		Variables Included: Fundamentals + ….						
	Fundamentals	Occupation	Goods	Shelter	Family Life	Soft Info App	Soft Info Credit	All
Observations	307,321	307,321	307,045	307,321	307,321	307,321	306,302	306,026
Pseudo R-squared	**0.0872**	**0.0944**	**0.0937**	**0.0885**	**0.0872**	**0.0916**	**0.0904**	**0.108**
Area under ROC	**0.7217**	**0.7297**	**0.7289**	**0.7232**	**0.7217**	**0.7262**	**0.7255**	**0.7434**

Step 2: Which of those Proxy Variables pass the Input Accountability Test?

Example: test the variable “elevators”.

- First, start with linear Decomposition: Proxy = fundamentals + residual
- Second: test if residual is correlated with female

Regress: $\text{Elevators} = a_1*\text{creditscore}+a_1*\text{income}+a_2*\text{debt}+\ldots.a_N*\text{lastFundamantal}+ \text{residual}$

Regress: $\text{Residual} = b0 + b1* \text{female}$

Test: $b1 \mathrel{!}= 0$

- Concern: p-value on b1…. decreases with the number of observations mechanically

- Cannot go down an “economic significance” argument because this is law. There is no sense in the law that “5 people out of 10,000 do not matter”

- d-value approach to the p-value problem as n-> large

D-value : Demidenko (2013)

“The P-value You Can’t Buy” American Statistician

- Rather than focus on a comparison of group means, the d-value is designed to examine how a randomly chosen female fared under this proxy variable relative to a randomly chosen male.

P value (under normality):

$$p = \Phi\left(-\frac{|b|}{s}\right)$$

D-value (under normality):

$$d = \Phi\left(-\frac{|b|}{s\sqrt{n}}\right)$$

Where s is the standard error: $s = \text{stdev}/\sqrt{n}$

<u>Foundations:</u>

- Individual observation comparison of this form are the foundation of the Wilcoxon-Mann-Whitney U Stat (for medians test)
- “D” comes from “discrimination” because the formulation is the same as the area under the ROC curve used for discrimination tests as early as Bamber (1975)

Family Lifestyle

	(1) Civil Marriage	(2) Non-civil Marriage	(3) Widow	(4) # Children	(5) Rural	(6) Large Metro
Coefficient from logit (default)	not signif.	-0.0999***	-0.146***	not signif.	-0.198***	0.0915***
Sign on residual estimation below that would indicate algorithmic bias against females	none	–	–	none	–	+
		Regression:	**Residual = b0 + b1* female**			
female	0.0174	-0.0684	0.042	-0.00596	0.0112	0.00604
	[0.00112]	[0.00177]	[0.000833]	[0.00272]	[0.00110]	[0.00136]
Observations	307,321	307,321	307,321	307,321	307,321	307,321
R-squared	0.001	0.005	0.008	0.000	0.000	0.000
Standard errors in brackets						
On d-values below: range +/- 1% around 50% is not concerning						
d-value		47.2%	53.6%		50.7%	50.3%

Family Lifestyle

Having a non-civil marriage lowers default risk. Thus the scoring algorithm rewards those of this category.

	(1)	(2)	(3)	(4)	(5)	(6)
	Civil Marriage	Non-civil Marriage	Wi			Metro
Coefficient from logit (default)	not signif.	-0.0999***	-0.14			5***
Sign on residual estimation below that would indicate algorithmic bias against females	none	–	-			

Regression: Residual = b0 + b1* female

	(1)	(2)	(3)	(4)	(5)	(6)
female	0.0174	-0.0684	0.042	-0.00596	0.0112	0.00604
	[0.00112]	[0.00177]	[0.000833]	[0.00272]	[0.00110]	[0.00136]
Observations	307,321	307,321	307,321	307,321	307,321	307,321
R-squared	0.001	0.005	0			00

Standard errors in brackets

On d-values below: range +/- 1% around 50% is not concerning

d-value		47.2%	53.6%		50.7%	50.3%

But the residual of non-civil marriage after orthogonalizing to the credit risk fundamentals is negatively correlated with being female. Thus the use of this variable overly penalizes females.

Is it significant? Yes. The d-value is different from 50% by >1%

Eliminate Results across all categories

Eliminated ONLY **3 of 37 variables** for bias

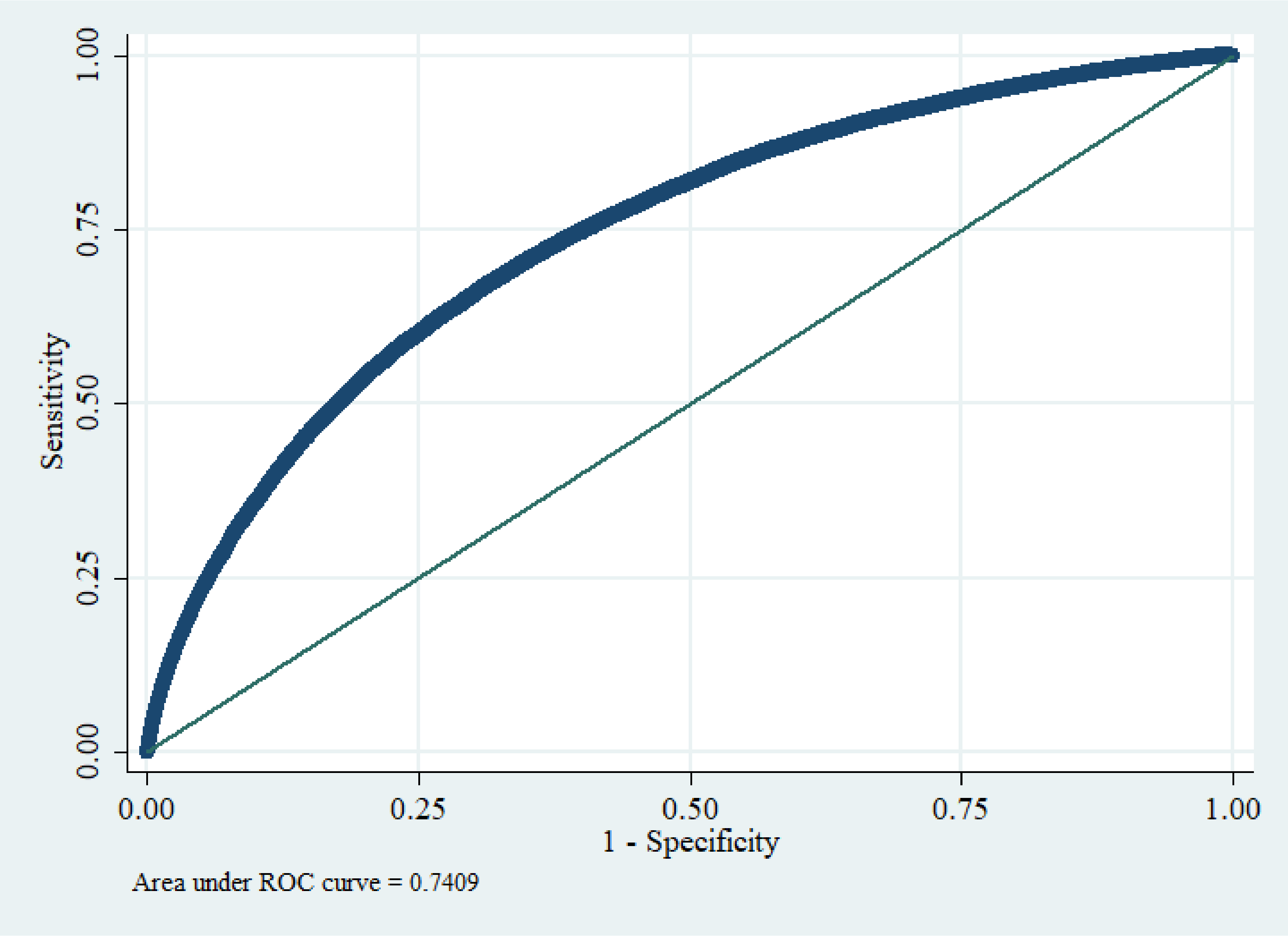
- previous goods loan-to-value
- non-civil marriage
- gives phone number for employer

How much area under the ROC curve / pseudo r-square is sacrificed?

Re-running Logit (default) dropping biased proxies

Area under ROC drops from 0.7434 to 0.7409

Pseudo rsquared drops from 0.108 to 0.1054



To do’s

1. What is the cost in dollars and counts of people from a wrong prediction due to excluding the variables failing the IAT?
2. What if one does not have all the fundamental variables?
   - Step into the benefit of each grouping of variables
   - Then the cost of failing the IAT is more, presumably
3. Add in the final dataset of credit card transaction data
4. Interactions? More ML?

Conclusions

Objectives:

- Get more finance research engaged in the policy debate about algorithmic use in credit scoring
- Debunk the emerging literature that AI poses no danger because it removes discretion, and any biases can be corrected

Accomplished (hopefully)

1) Demonstrated what the law dictates about inputs & business necessity
2) Provided a really simple test for firms to use ex ante and regulators or courts ex post
3) Showed that at least in our application, the test provides results that are workable to firms