Predictably Unequal?
The Effects of Machine Learning on Credit Markets

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, Ansgar Walther

SNB, Yale SOM, Imperial College ($\times$2)

September 2020

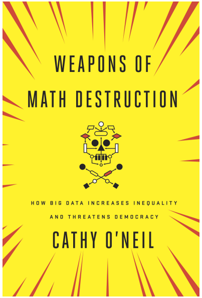
Advances in Technology and Inequality

- Machine learning has been rapidly adopted in many industries
- Central application: default prediction in credit markets
  (e.g. Khandani, Kim, and Lo, 2010; Sirignano, Sadhwani, and Giesecke, 2017)
- **This paper: What are the distributional effects of new technology?**

Advances in Technology and Inequality

Capital One

vs.



This Paper

Theory: Distributional implications of “better” statistical technology

Mortgage default prediction: Using US administrative data with traditional technology (Logit) and Machine Learning

Distributional consequences of new technology

- Across racial groups: fewer winners in some minority groups; increased dispersion

Equilibrium implications in a model of competitive loan pricing

- Outcomes differ on both extensive and intensive margins

A Lender's Prediction Problem

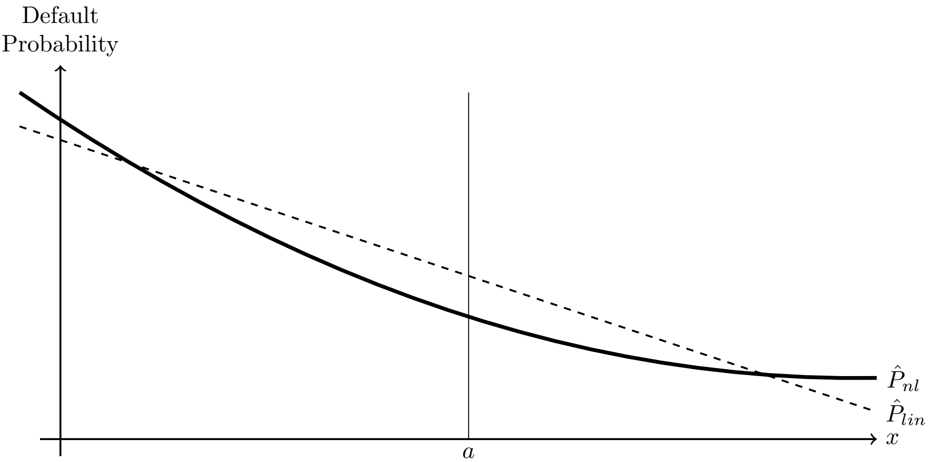
Observe borrowers with characteristics x and default outcome y

Predict $\hat{y} = \hat{P}(x)$ to minimize MSE

- Old technology: Restricted class of functions $\hat{P}$ (e.g. linear)
- New technology: Wider class of permitted functions

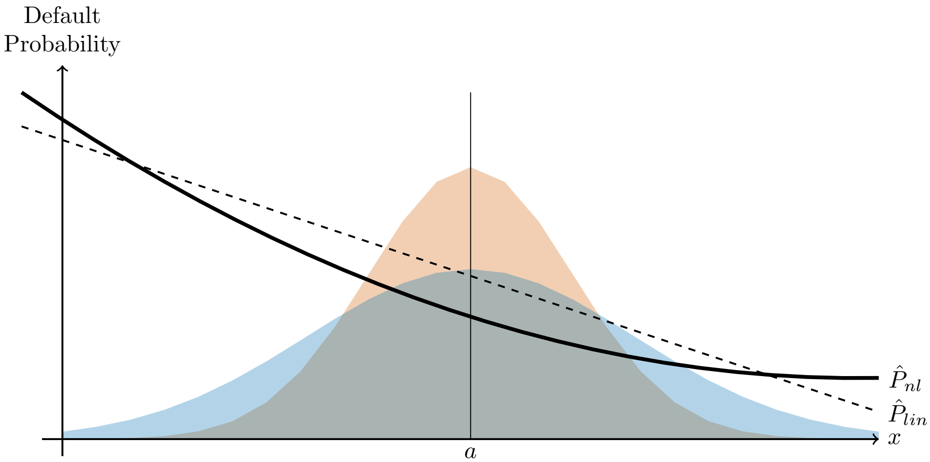
Lemma. Optimal predictions with new technology are a *mean-preserving spread* of those with old technology $\Rightarrow$ *There are winners and losers*

Winners and Losers



Convex quadratic: “extreme” x lose, others gain

Winners and Losers



Two groups: "blue" borrowers lose due to high variance

Sources of Unequal Effects

- Previous example could arise from

$$y = P(x) + \varepsilon,$$

where P is nonlinear and the group g does not matter for y.

$\Rightarrow$ Winners/losers arise from additional **flexibility** of new technology.

Effects across g depend on functional form of new technology, and the differences in distribution of characteristics

Sources of Unequal Effects

- Previous example could arise from

$$y = P(x) + \varepsilon,$$

  where P is nonlinear and the group g does not matter for y.

  $\Rightarrow$ Winners/losers arise from additional **flexibility** of new technology.

  Effects across g depend on functional form of new technology, and the differences in distribution of characteristics

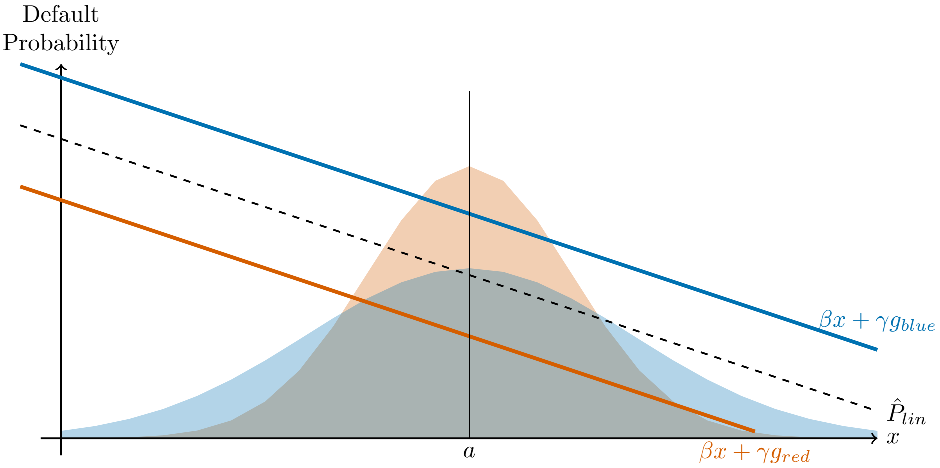
- Alternative:

$$y = \beta \cdot x + \gamma \cdot g + \varepsilon,$$

  i.e. true relationship is linear, but g predictive of default.

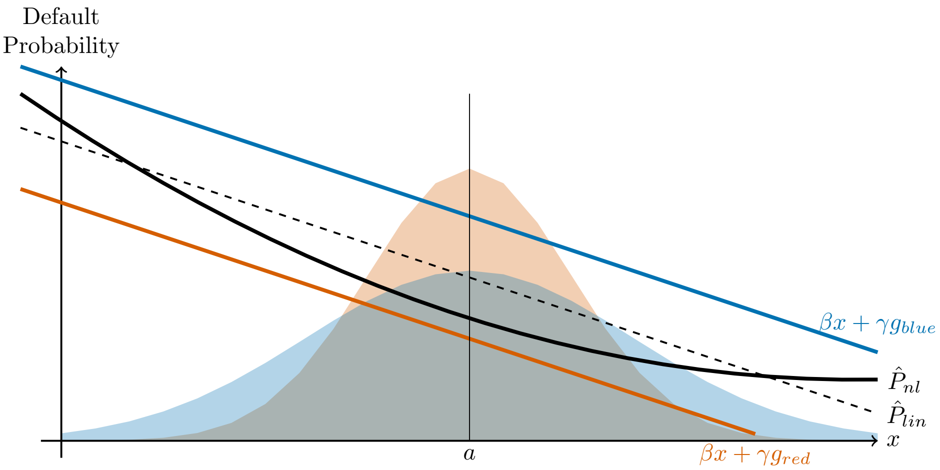
  $\Rightarrow$ Effects of new technology arise due to "**triangulating**" g

Triangulation



- No linear correlation between x and g $\rightarrow$ linear model simply recovers average

Triangulation



- Blue borrowers more likely to have extreme $x \rightarrow$ nonlinear model penalizes.

US Mortgage Data

HMDA

- Application date, applicant income, loan type, size, purpose,
- **race, ethnicity, gender**

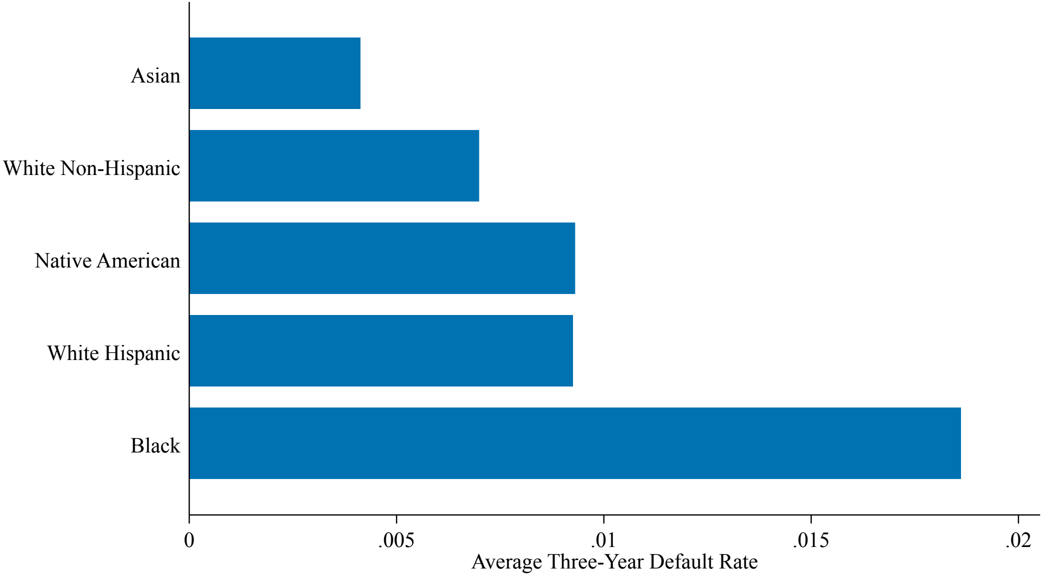
McDash (Black Knight)

- Underwriting, contract and performance: e.g. FICO, LTV, interest rate, **default status**

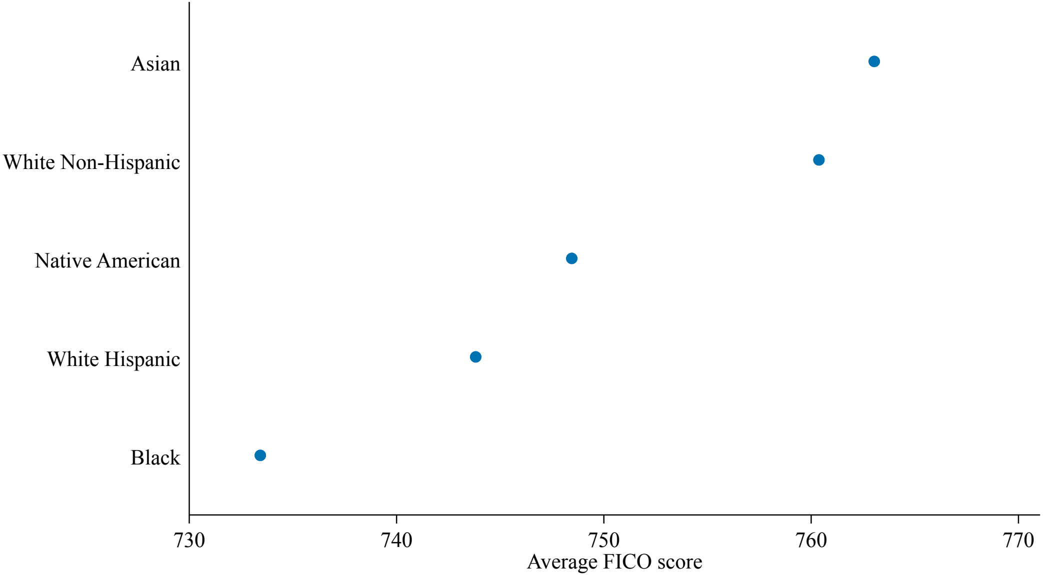
Linked Dataset

- 9.4m mortgage loans from 2009-2013
- Portfolio *and* GSE loans, $<$ \$1m
- **Default**: 90+ days delinquent within 3 years of origination

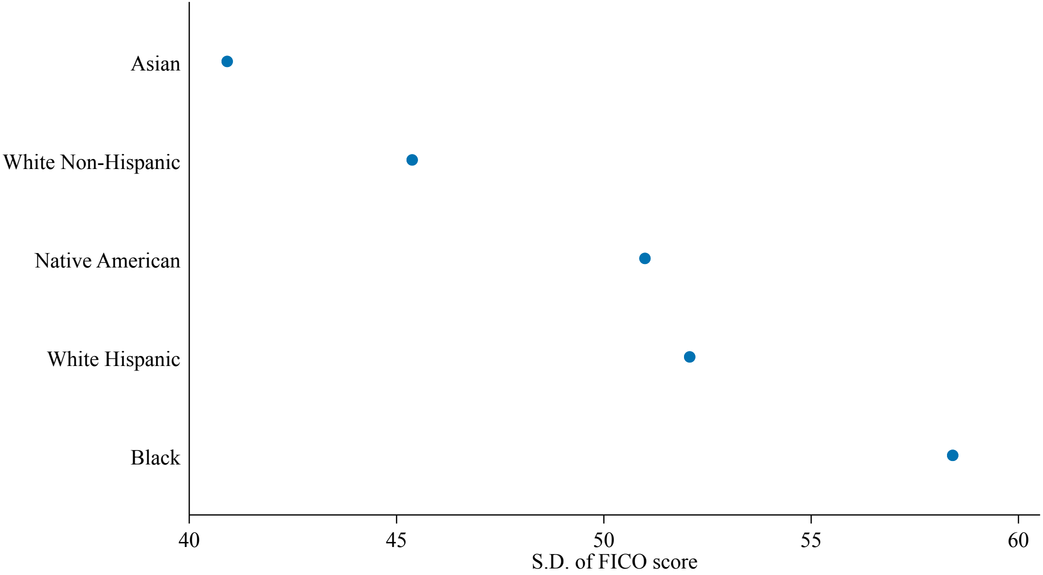
Default Rates Across Race



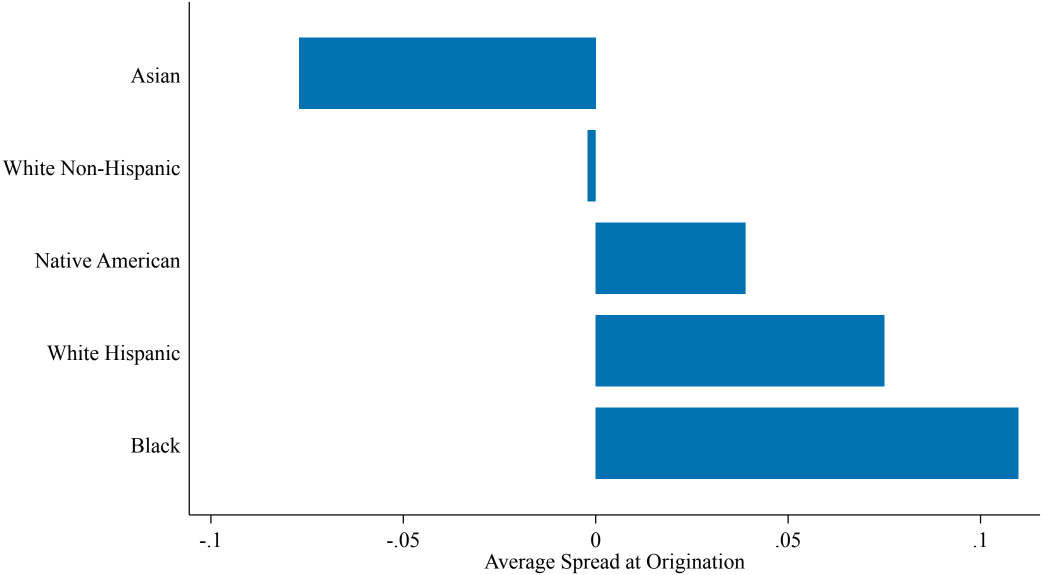
Mean FICO Across Race



S.D. of FICO Across Race



Interest Rates Across Race

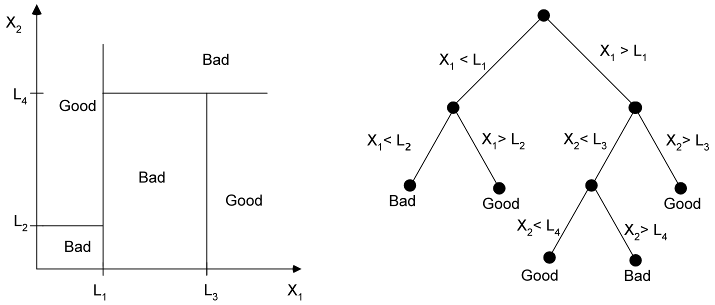


Estimating Probabilities of Default: Technologies

Traditional: Probability of Default = Logit(x) (e.g. Demyanyk and Van Hemert, 2011; Elul et al., 2010)

- Using nonlinear "bin" dummies for FICO, LTV, income

Machine Learning: Decision trees estimate step functions



(from Khandani, Kim, and Lo, 2010)

1. Random forest (w/cross-validation)
2. Calibration (isotonic regression)

- (Similar if use "XGBoost")

Explanatory Variables

Logit	*Nonlinear Logit*
Applicant Income (linear)	Applicant Income (25k bins, from 0-500k)
LTV Ratio (linear)	LTV Ratio (5-point bins, from 20 to 100%; separate dummy for LTV=80%)
FICO (linear)	FICO (20-point bins, from 600 to 850;) separate dummy for FICO$<$600)
(with dummy variables for missing values)	
Common Covariates	
Spread at Origination "SATO" (linear)	
Origination Amount (linear and log)	
Documentation Type (dummies for full/low/no/unknown documentation)	
Occupancy Type (dummies for vacation/investment property)	
Jumbo Loan (dummy)	
Coapplicant Present (dummy)	
Loan Purpose (dummies for purchase, refinance, home improvement)	
Loan Term (dummies for 10, 15, 20, 30 year terms)	
Funding Source (dummies for portfolio, Fannie Mae, Freddie Mac, other)	
Mortgage Insurance (dummy)	
State (dummies)	
Year of Origination (dummies)	

Model Performance ‣ More Detail

Estimate on *training set* (70%), evaluate on *test set* (30%).

Out-of-sample performance:

- $R^2 \uparrow$ by 14.30%
- *Precision Score* $\uparrow$ by 5.1%

  How many predicted defaults are true defaults?

- Bootstrap analysis confirms significant differences

$\rightarrow$ Random Forest method substantially better predictor of $\Pr(\mathit{default}|X)$

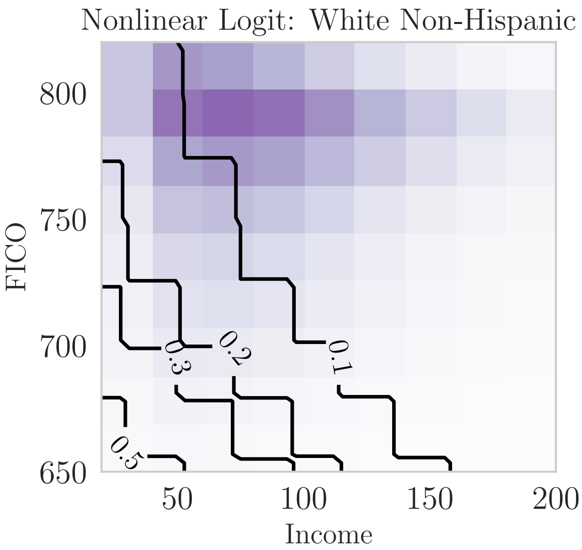
Unequal Effects of New Technology: Example

- Fix all characteristics but income + FICO

- Compare distribution vs. predictions by race

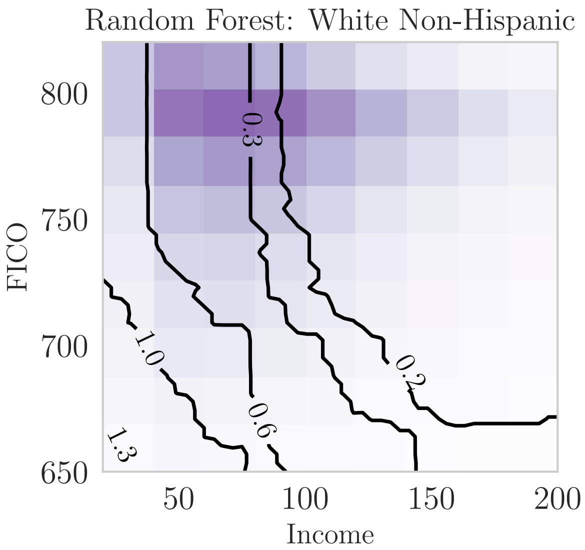
Unequal Effects of New Technology: Example

- Fix all characteristics but income + FICO
- Compare distribution vs. predictions by race
- Logit not very flexible



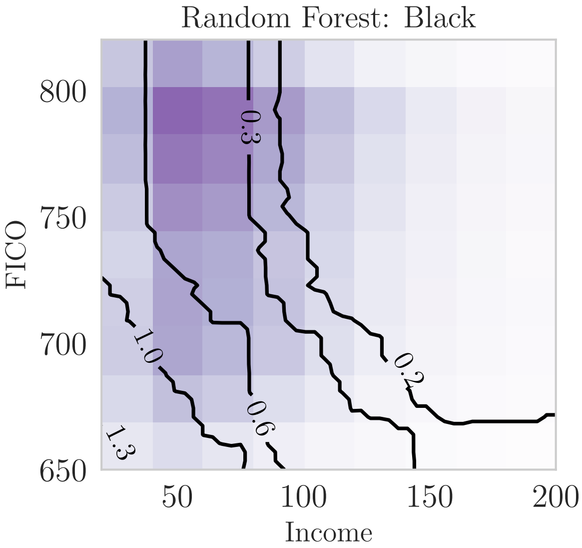
Unequal Effects of New Technology: Example

- Fix all characteristics but income + FICO
- Compare distribution vs. predictions by race
- Logit not very flexible
- RF much more flexible

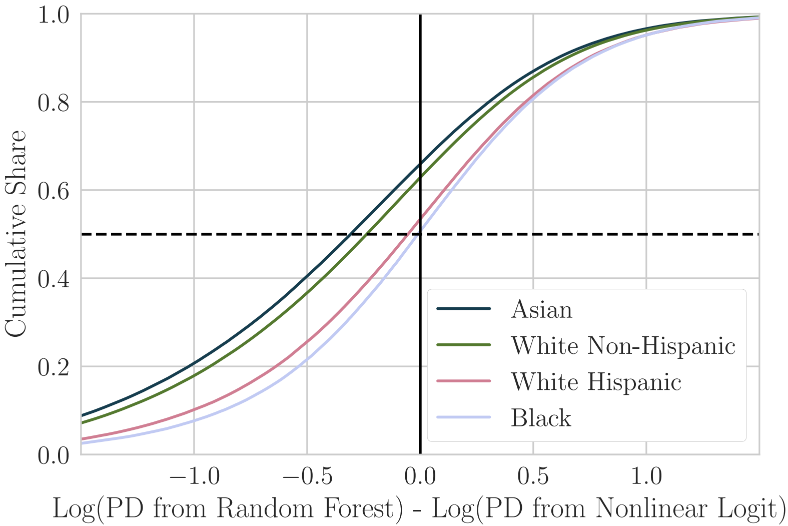


Unequal Effects of New Technology: Example

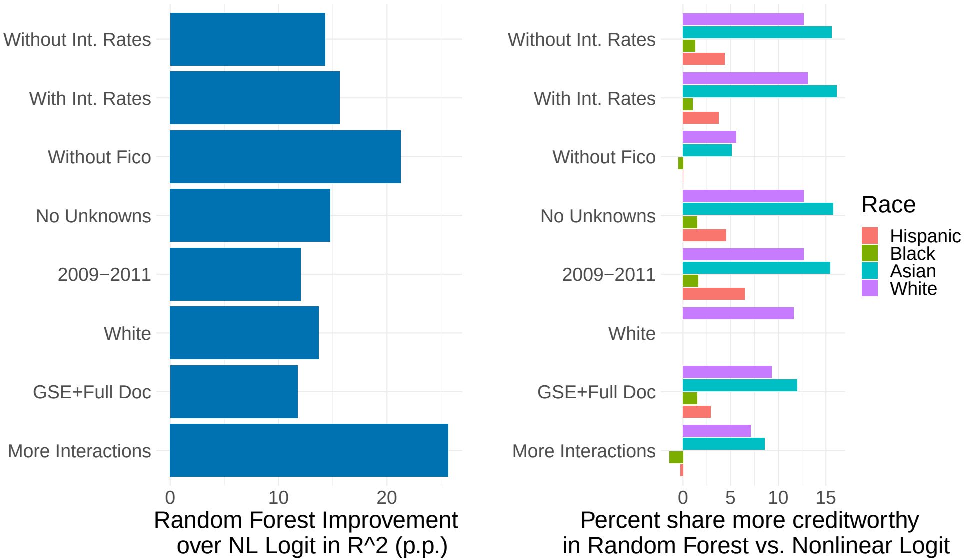
- Fix all characteristics but income + FICO
- Compare distribution vs. predictions by race
- Logit not very flexible
- RF much more flexible



Unequal Effects of New Technology: Population



Unequal Effects of New Technology: Alternative Approaches



Predicting Race using Identical Observables

Model	ROC AUC	Precision Score	Brier Score $\times$ 10	R^2
Logit	0.7478	0.1948	0.5791	0.0609
Nonlinear Logit	0.7485	0.1974	0.5783	0.0622
Random Forest	0.7527	0.2110	0.5665	0.0813

$\longrightarrow$ RF model is strikingly better at predicting black / hispanic borrowers

Flexibility versus Triangulation

Decomposition of model improvements:

1. Add **race** as an explanatory variable to Logit
2. Allow use of ML **technology** to the model with race
   (i.e. ”add” nonlinear functions / interactions of x as explanatory variables)

Flexibility versus Triangulation

Decomposition of model improvements:

1. Add **race** as an explanatory variable to Logit
2. Allow use of ML **technology** to the model with race
   (i.e. ”add” nonlinear functions / interactions of x as explanatory variables)

	Race	Race Int.	Technology
ROC-AUC	6.28	2.04	91.69
Precision	9.05	22.43	68.52
R2	3.37	4.39	92.24

$\Rightarrow$ Improved performance mostly due to flexibility, not triangulation

Flexibility versus Triangulation

Decomposition of model improvements:

1. Add **race** as an explanatory variable to Logit
2. Allow use of ML **technology** to the model with race
   (i.e. "add" nonlinear functions / interactions of x as explanatory variables)

	Technology	Race
ROC-AUC	89.77	10.23
Precision	94.14	5.86
R2	92.95	7.05

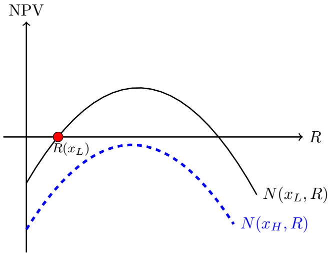
$\Rightarrow$ Improved performance mostly due to flexibility, not triangulation

Interest Rates in Competitive Equilibrium

Simple 2-period model:

$$NPV(x,R) = \frac{1}{1+\rho}\left[(1-P(x,R))(1+R)L + P(x,R)\tilde{L}\right] - L$$

- Equilibrium $R^\star(x)$ solves $NPV = 0$
- Reject x-borrowers if $NPV(x,R) < 0$ for all feasible R

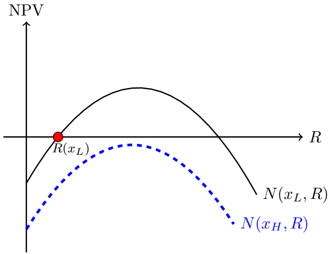


Interest Rates in Competitive Equilibrium

Simple 2-period model:

$$NPV(x,R) = \frac{1}{1+\rho}\left[(1-P(x,R))(1+R)L + P(x,R)\tilde{L}\right] - L$$

- Equilibrium $R^\star(x)$ solves $NPV = 0$
- Reject x-borrowers if $NPV(x,R) < 0$ for all feasible R



- Calibration:
  - recovery: $\tilde{L} = min((1+R)L, 0.75V) - 0.1L$ (second part: carrying costs, liquidation exp.)
  - WACC: ρ = quarterly average interest rate -30bps
  - 3-year PD to lifetime via "standard default assumption" (MBS mkt convention)

Identification

In data, only observe one R per loan. But likely not randomly allocated.
$\rightarrow$ bias in $\hat{P}$ for counterfactual rates R

Proposed solution:

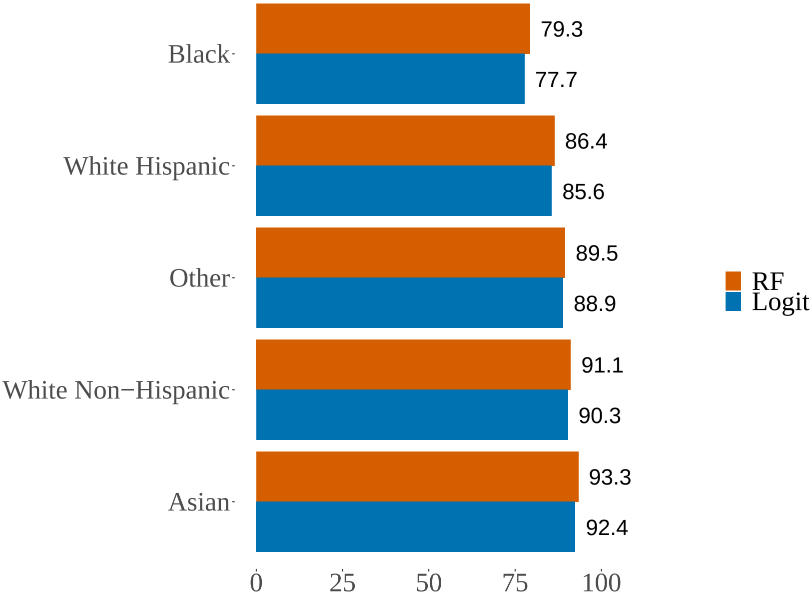
1. Restrict attention to GSE / full documentation loans
   $\rightarrow$ Likely selection on *observable* variables, not soft information (Keys et al., 2010)

2. Adjust $\frac{\partial \hat{P}}{\partial R}$ downwards using ratio of causal to reduced-form estimates based on Fuster and Willen (2017)

   - Estimated $\frac{\partial \hat{P}}{\partial R}$ over first 3 years $\approx 1.7\times$ causal $\frac{\partial P}{\partial R}$

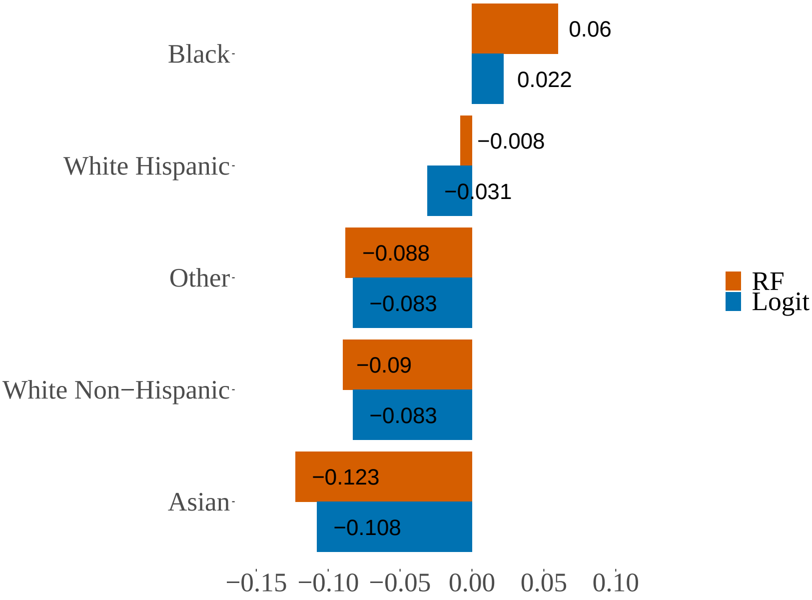
Model Outcomes

- Acceptance rates



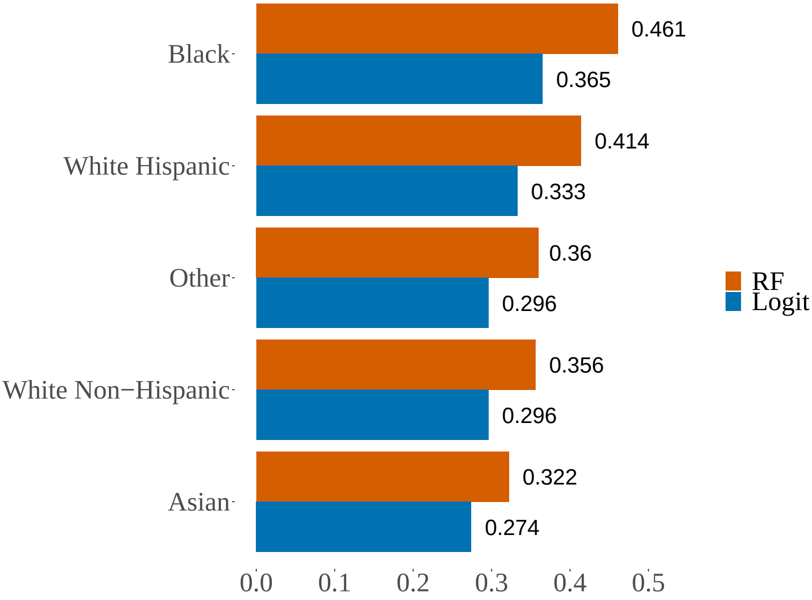
Model Outcomes

- Acceptance rates
- Average SATO ($= R - \bar{R}_t$)



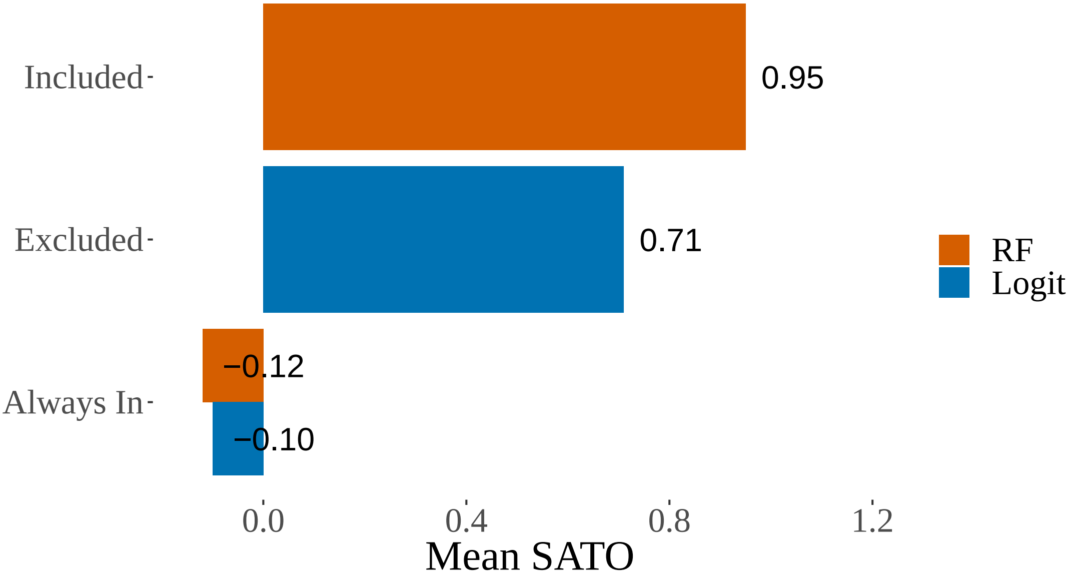
Model Outcomes

- Acceptance rates
- Average SATO ($= R - \bar{R}_t$)
- S.D. of SATO $\rightarrow$ new technology increases dispersion *across* and *within* groups



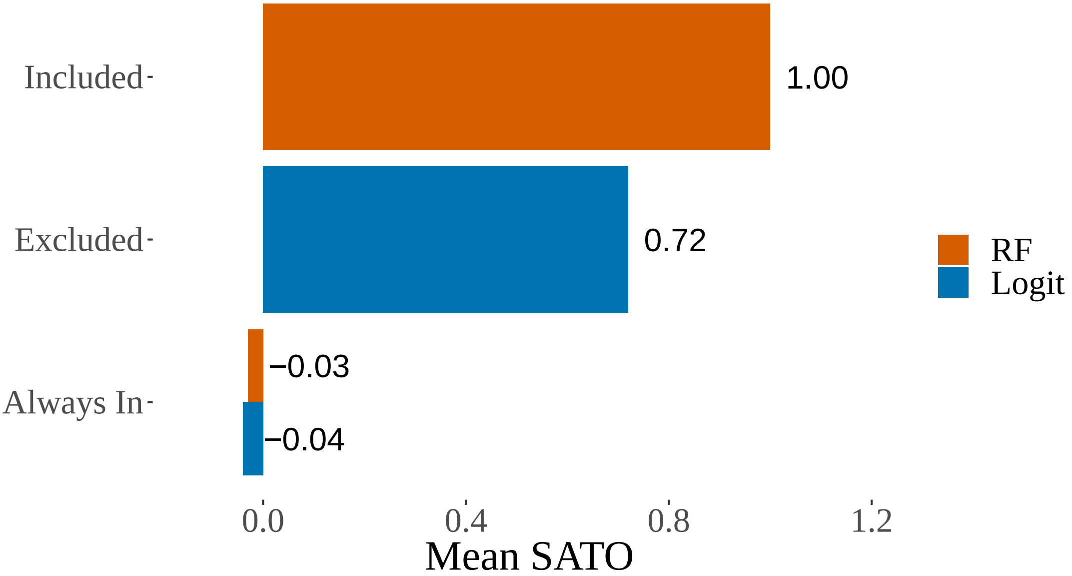
Inclusion and Exclusion in Equilibrium

Average SATO: White + Asian Borrowers



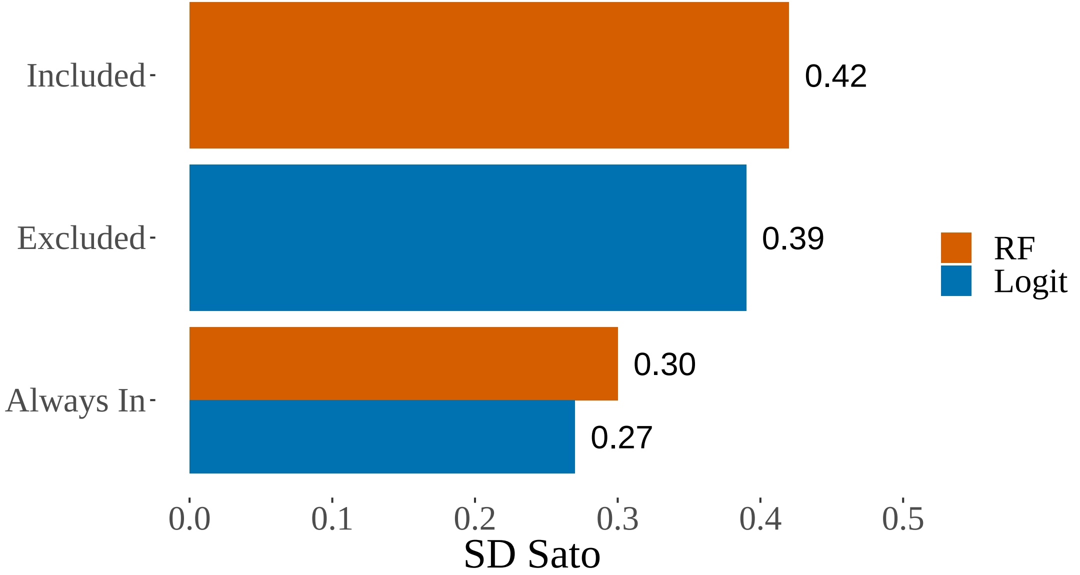
Inclusion and Exclusion in Equilibrium

Average SATO: Black + Hispanic Borrowers



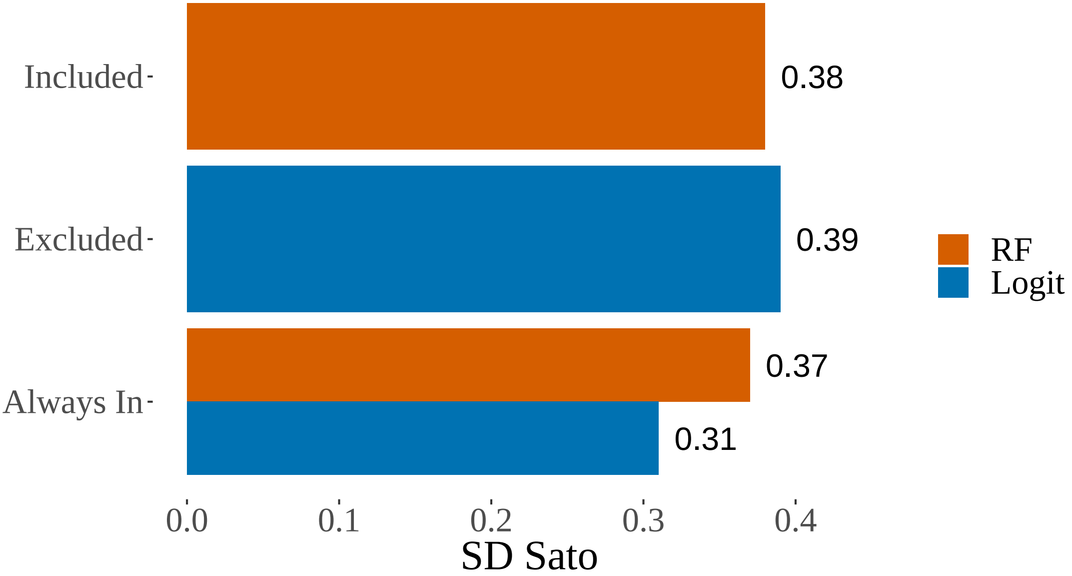
Inclusion and Exclusion in Equilibrium

SD of SATO: White + Asian Borrowers



Inclusion and Exclusion in Equilibrium

SD of SATO: Black + Hispanic Borrowers



Conclusion

- Improvements in statistical technology creates
  - Greater predictive power and gains for producers
  - Increased disparity in outcomes for consumers
- Based on US mortgage data, black + hispanic borrowers bear larger changes
  - First-moment effects: More likely to be perceived as high risk
  - Second-moment effects: Greater increase in dispersion of outcomes
  - Improvement comes from more than just “putting race in”
- Equilibrium effects
  - Positive extensive-margin effect of new technology
  - Unequal effects persist at intensive margin

Performance of Different Statistical Technologies Predicting Default

	ROC AUC		Precision Score		Brier Score $\times$ 100		R^2	
Model	(1) No Race	(2) Race	(3) No Race	(4) Race	(5) No Race	(6) Race	(7) No Race	(8) Race
Logit	0.8522	0.8526	0.0589	0.0592	0.7172	0.7171	0.0245	0.0246
Nonlinear Logit	0.8569	0.8573	0.0598	0.0601	0.7146	0.7145	0.0280	0.0281
Random Forest	0.8634	0.8641	0.0630	0.0641	0.7114	0.7110	0.0323	0.0329

Measuring Model Performance

▸ Return

MSE is one natural way to evaluate $\hat{P}$:

$$MSE(\hat{P}) = n^{-1} \sum_{n} (\hat{P}(x_i) - y_i)^2$$

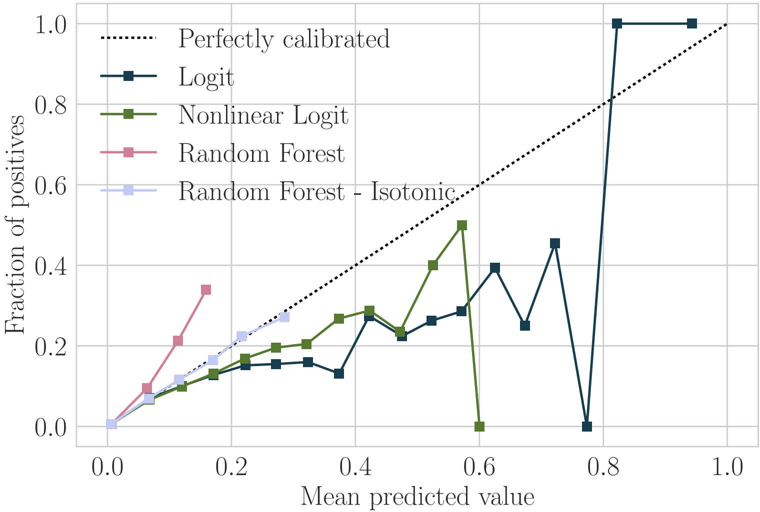
Can be decomposed into three components:

$$MSE(\hat{P}) = \underbrace{n^{-1} \sum_{k=1}^{K} n_k (\hat{y}_k - \bar{y}_k)^2}_{\text{Reliability}} - \underbrace{n^{-1} \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2}_{\text{Resolution}} + \underbrace{\bar{y}(1 - \bar{y})}_{\text{Uncertainty}}$$

- Brier Scores: RF 0.00711, Logit 0.00714, but overall uncertainty: 0.00735
- Reliability: Logit is 3500% worse than RF
- Resolution: Logit is 40% better than RF

Isotonic regressions and calibration

▸ Return



Decomposition of Performance Improvement

▸ Return

	Race	Technology
ROC-AUC	5.88	94.12
Precision	7.90	92.10
Brier	3.25	96.75
R^2	2.04	97.96

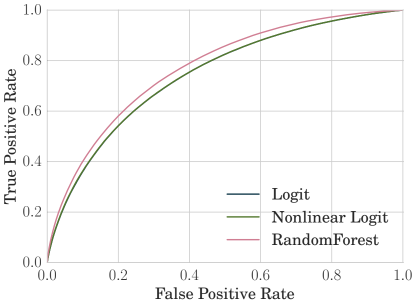
Panel A: Race Controls First

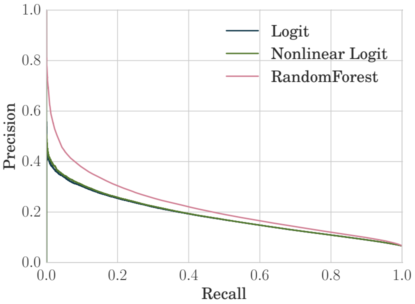
	Technology	Race
ROC-AUC	91.16	8.84
Precision	77.21	22.79
Brier	90.63	9.37
R^2	87.75	12.25

Panel B: New Technology First

- Panel A: Nonlinear Logit $\rightarrow$ add race dummies. Get less than 8% of fit improvement that get from moving to Random Forest (w/o race)
- Panel B: Random Forest $\rightarrow$ add race dummies. Slightly larger improvements from having race but still much less important than benefit of flexibility

Predicting Minority Status from x ▸ Return





Model	ROC AUC	Precision Score	Brier Score $\times$ 10	R^2
Logit	0.7478	0.1948	0.5791	0.0609
Nonlinear Logit	0.7485	0.1974	0.5783	0.0622
Random Forest	0.7527	0.2110	0.5665	0.0813