Algorithmic Stereotyping:
Overview and Key Considerations for More Ethical AI

Val Carey 11/12/2020

Algorithmic Bias

“Systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others” (Wikipedia)

The Apple Card

DHH @dhh · Nov 7, 2019

The @AppleCard is such a bleep sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

1.4K 12.6K 28.2K

DHH @dhh · Nov 7, 2019

I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

83 266 4.3K

DHH @dhh · Nov 7, 2019

It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the bleep balance in advance and in full.

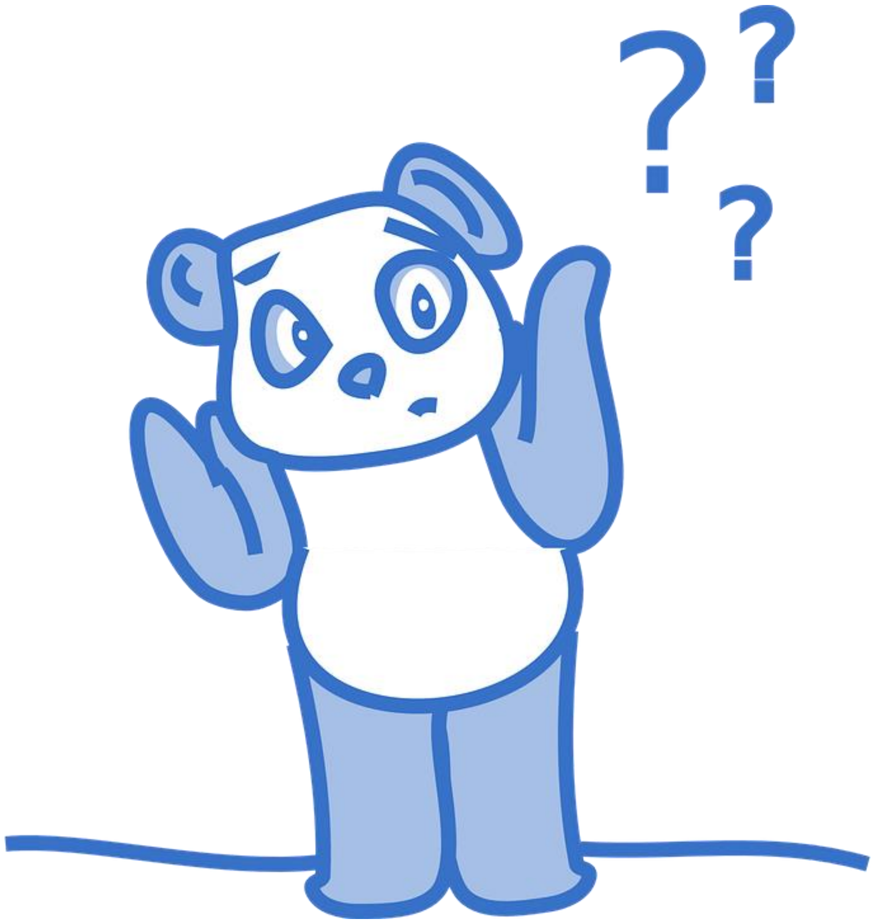
PAYCHEX
Payroll | Benefits | HR | Insurance

What Happened?

No one from [Apple] seemed able to describe how the algorithm even worked, let alone justify its output.

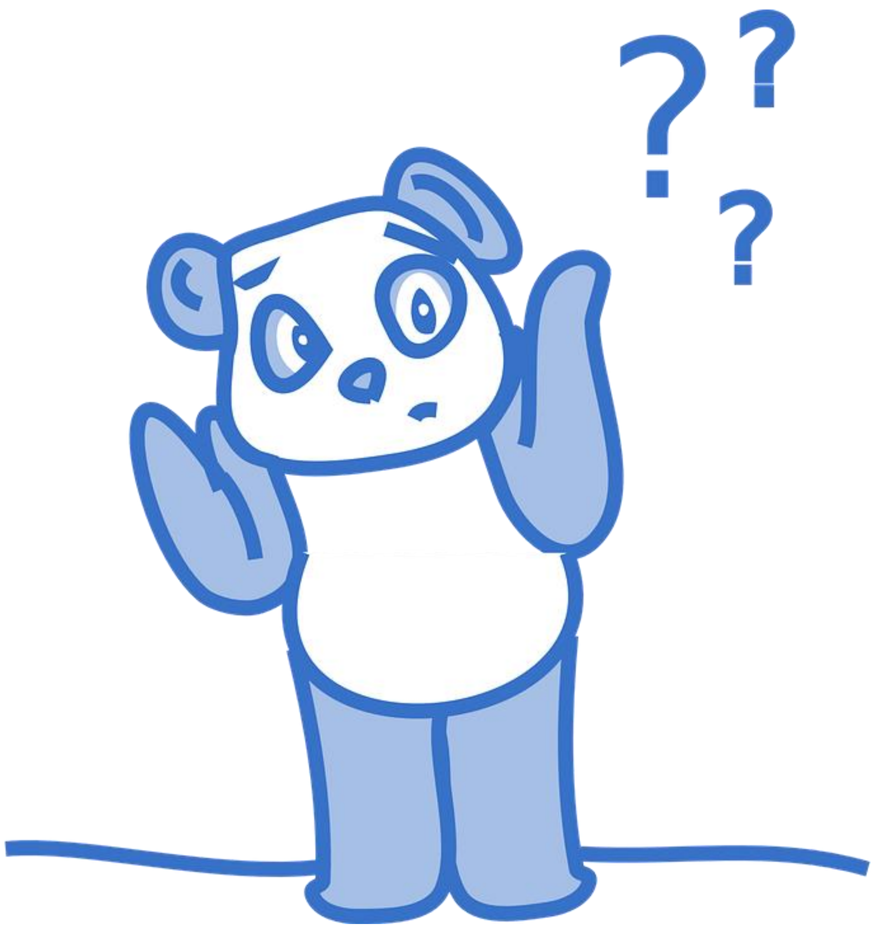
The algorithm …. doesn't even use gender as an input. How could the bank discriminate if no one ever tells it which customers are women and which are men?

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/



What Happened?

Predictive Analytics World: “Shopping patterns” were a proxy for gender (???)



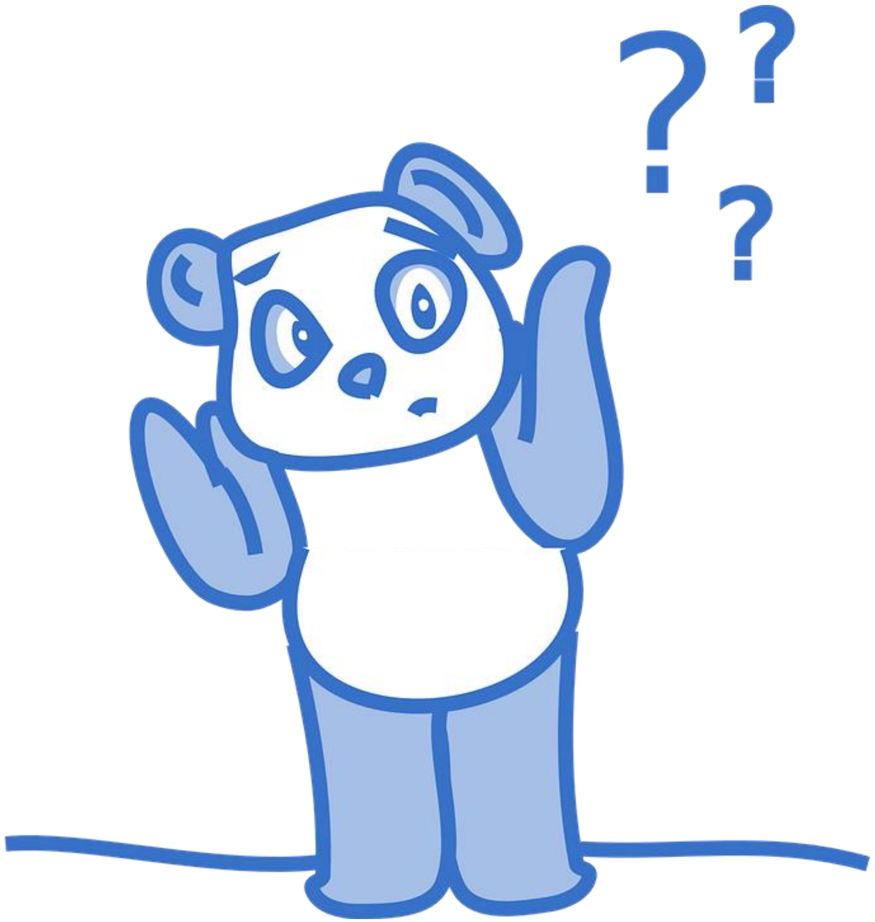
What Happened?

… the alleged discrimination might have less to do with the Apple Card and more to do with workplace discrimination, suggested Shayne Sherman, CEO of TechLoris.

"Women generally earn less than their male counterparts and are less likely to earn promotions….".

"This results in not only in current lower wages, but lower prospective wages, and ultimately lower credit limits…"

https://www.ecommercetimes.com/story/86351.html?__hstc=8228397.99a265337744294b740e0787aea508c4.1574294400195.1574294400196.1574294400197.1&__hssc=8228397.1.1574294400198&__hsfp=1895241284



What Do You Think?

Given that women earn less, what is fair?

i. We use income alone to determine credit limits. On average, women get lower limits.

ii. We use shopping patterns in our model, and these are a proxy for income. On average, women get lower limits

iii. We use shopping patterns in our model, and these are a proxy for gender. On average, women are less creditworthy, and so women wind up with lower limits.

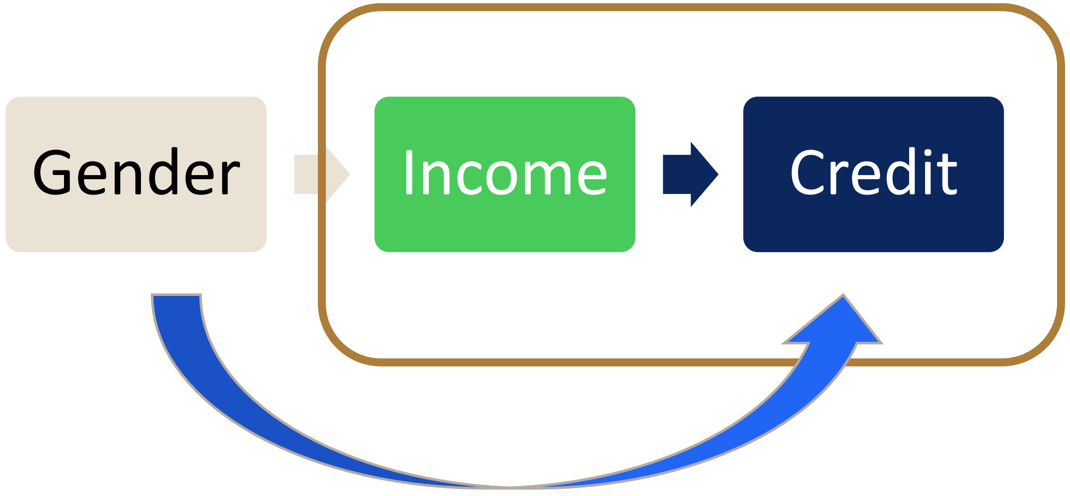
What Do You Think?

Given that women earn less, what is fair?

i. We use income alone to determine credit limits. On average, women get lower limits.

ii. We use shopping patterns in our model, and these are a proxy for income. On average, women get lower limits

iii. We use shopping patterns in our model, and these are a proxy for gender. On average, women are less creditworthy, and so women wind up with lower limits.



What Do You Think?

Given that women earn less, what is fair?

i. We use income alone to determine credit limits. On average, women get lower limits.

ii. We use shopping patterns in our model, and these are a proxy for income. On average, women get lower limits

iii. We use shopping patterns in our model, and these are a proxy for gender. On average, women are less creditworthy, and so women wind up with lower limits.

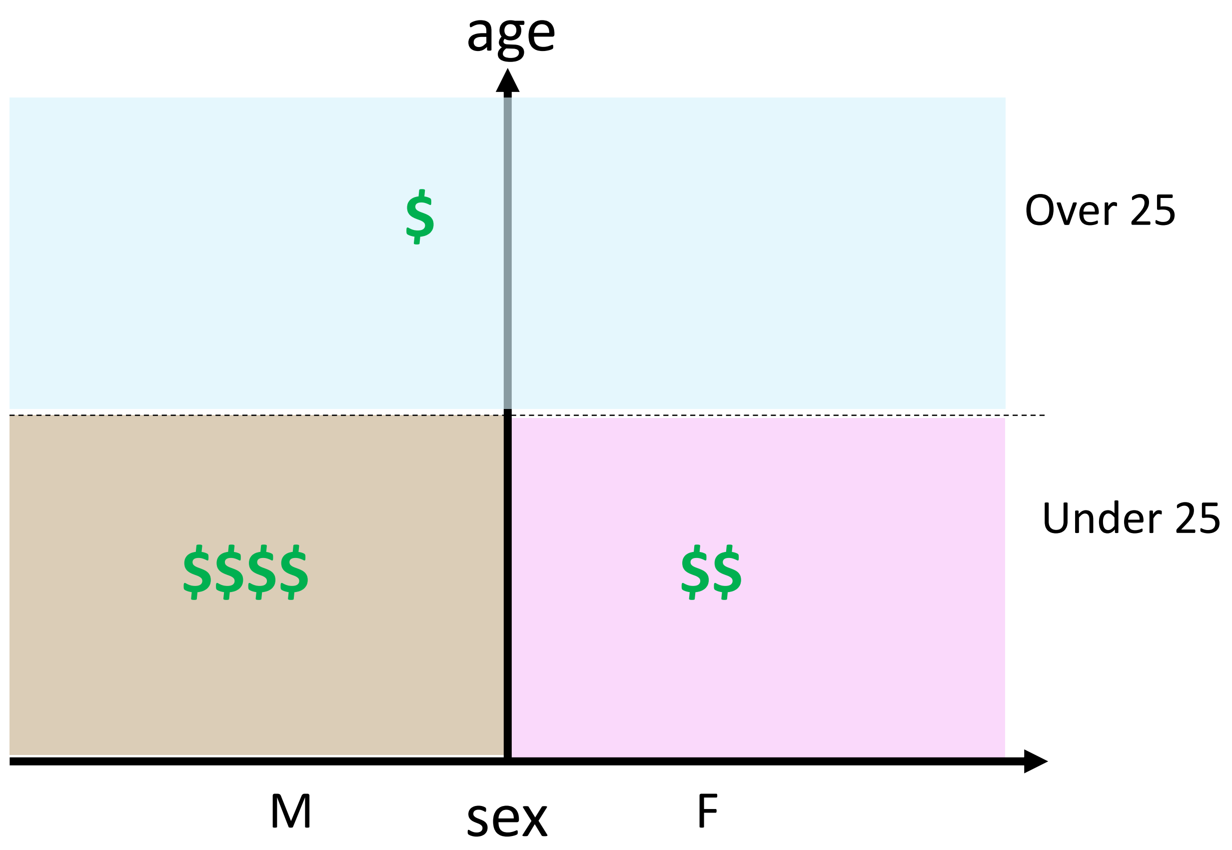
Bias : “Systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others” (Wikipedia)

PAYCHEX®
Payroll | Benefits | HR | Insurance

Actuarial Risk

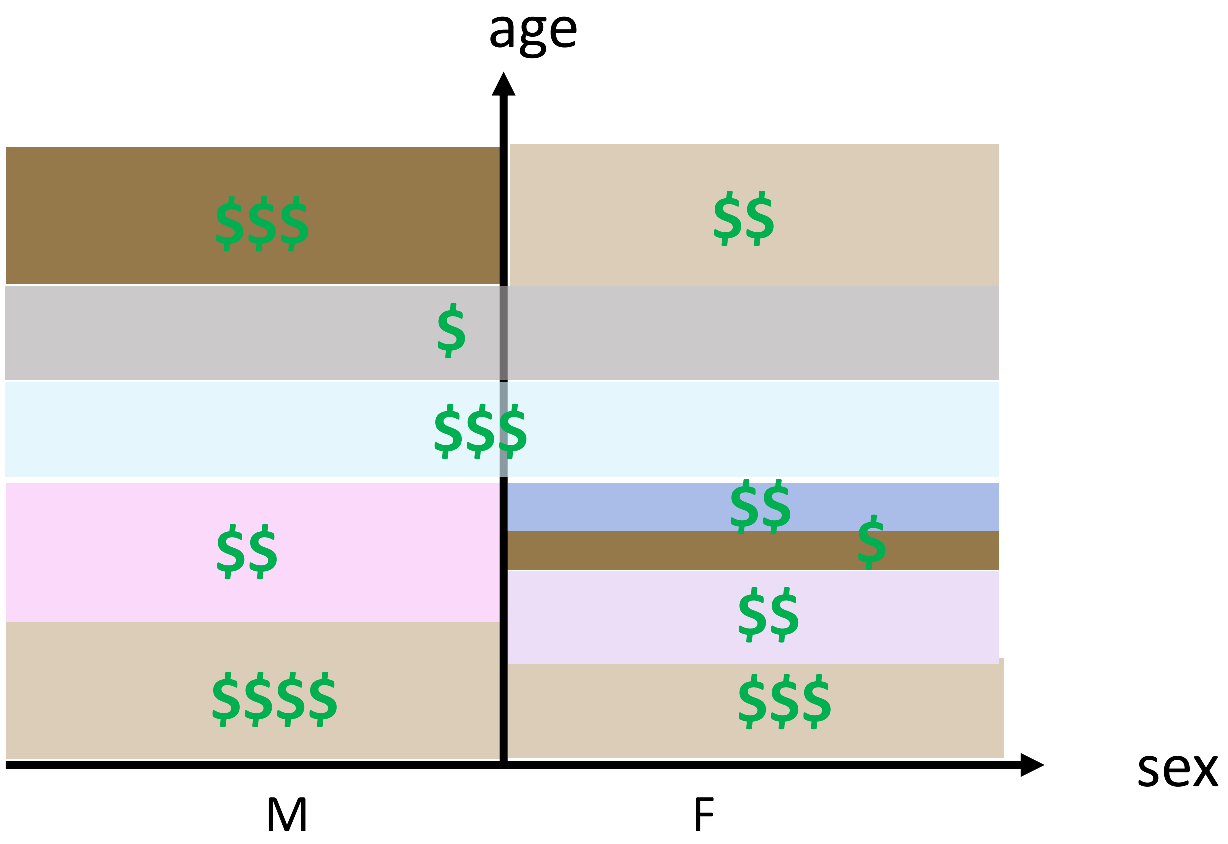
What does a model do?

What did they do before machine learning?

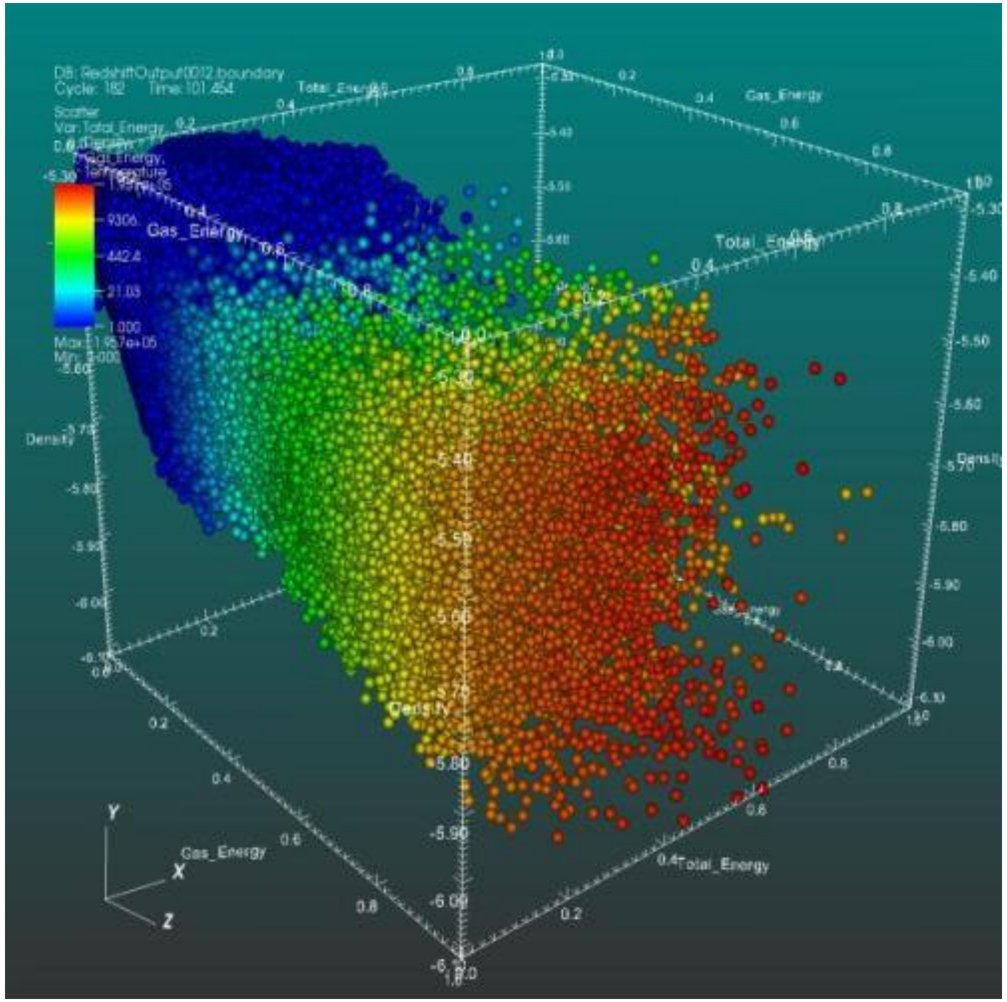


What does a model do?

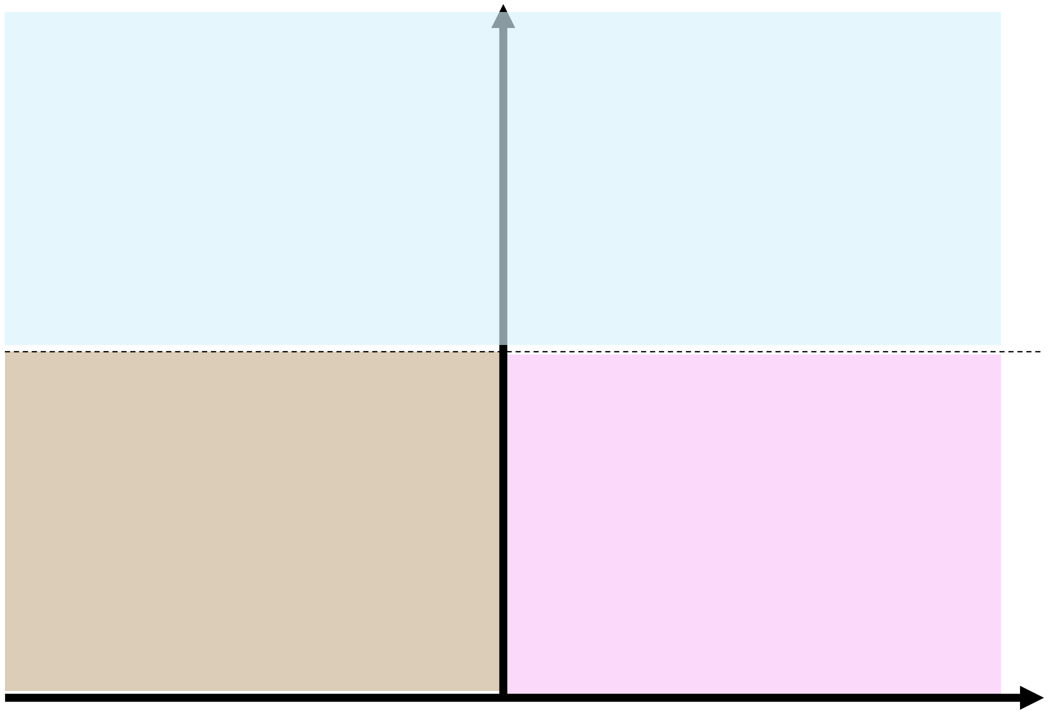
What do we do now?



What does a model do?



≈



What does a model do?

A model assigns a score based on the historical behavior of people **similar to you**

What does similar mean?

- Similar values of the features included in the model
- Features weighted according to the correlation with the response value

A model is a fancy actuarial table, which creates fine-grained groups and then labels them.

And then decisions are made.

"Who is like me, and what have they done before?"

"Out of 100 people with similar features, 10 will have a car accident in one year"

PAYCHEX®
Payroll | Benefits | HR | Insurance

There is nothing new under the sun

Or is there?

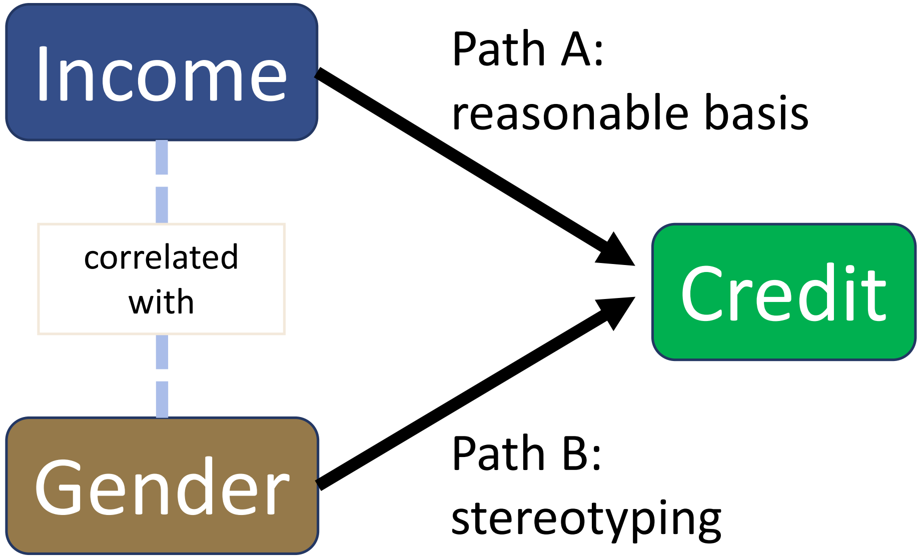
- Lack of transparency
- Volume and velocity
- Context

- Fairness metrics??

Fairness Metrics

"Apple Card" dilemma

If women on average have lower income, and income is correlated with loan repayment, is it fair to give women a lower score across the board, regardless of individual income?



Can "fairness metrics" distinguish Path A from Path B?

Demonstration

- Start with loans dataset with no gender information
- Assign “gender” to individuals based on income (probabilistically)
- Create model using original data but not “gender” (“Path A”)
- Create alternate model leaving out income and instead using gender (“Path B”)
- Calculate fairness metrics

“Lending club” data

163,987 rows, 13 predictive variables

https://raw.githubusercontent.com/h2oai/app-consumer-loan/master/data/loan.csv

1.	loan_amnt: num	(5000 2500 10000 3000 5375 ...)
2.	term: 2 levels	("36 months","60 months“)
3.	emp_length: num	(10 0 10 9 0 5 3 0 4 10)
4.	home_ownership: 6 levels	(“Rent”, “Own”, ”Mortgage”, …)
5.	annual_inc: num	(24000 30000 49200 48000 15000 72000 …)
6.	purpose: 14 levels	("car","credit_card",…)
7.	region: 4 levels	(“South”,”West”,”Northeast”,”Midwest”)
8.	dti: num	(27.65 1 20 5.35 18.08 ...)
9.	delinq_2yrs: num	(0 1 2 3 …)
10.	revol_util: num	(83.7 9.4 21 87.5 36.5 20.6 …)
11.	total_acc: num	(9 4 37 4 3 23 11 23 28 42 …)
12.	longest_credit_length: num	(26 12 15 4 7 13 8 4 13 18 …)
13.	verification_status: 2 levels	(“not verified", “verified”)
14.	bad_loan: 2 levels	("0","1“)

PAYCHEX®
Payroll | Benefits | HR | Insurance

Inferring “gender”

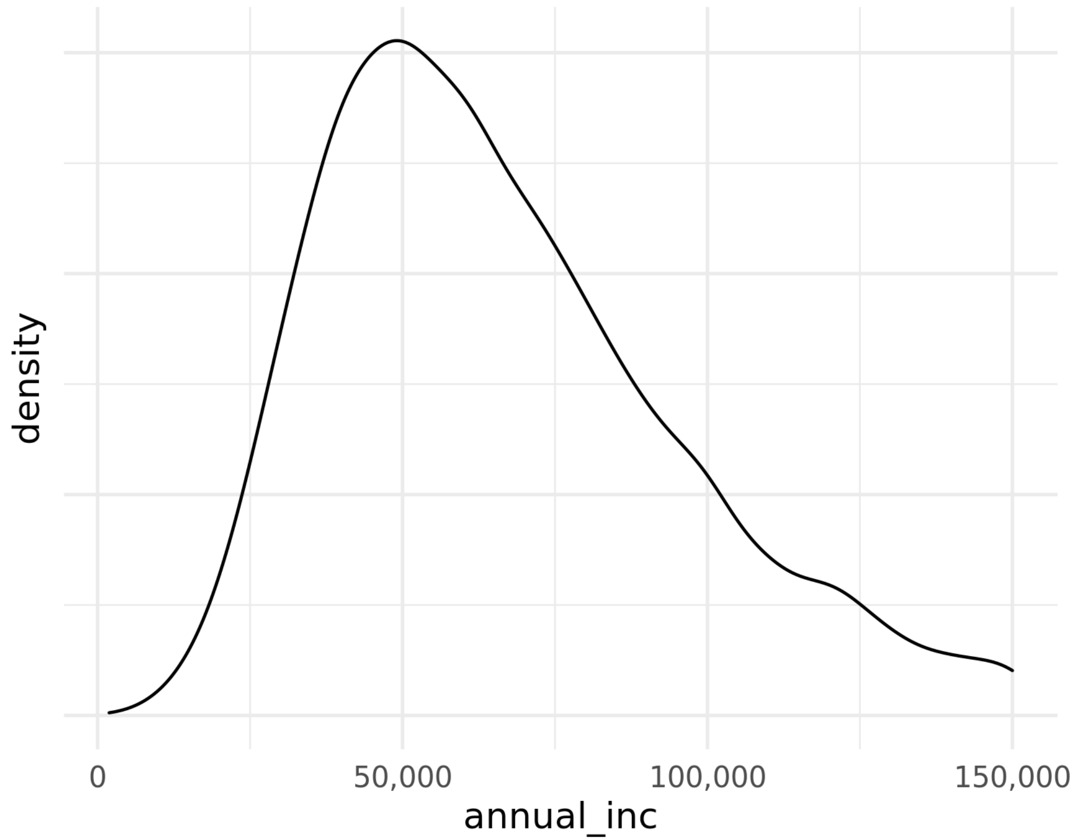
- Sigmoid function (algebraic form), midpoint at $45k,
- *This greatly exaggerates* sex-related differences in income (US)

inferred gender	count	percent	median income	loan default rate
F	41,724	26.4%	$38,000	21.9%
M	116,272	73.6%	$74,000	16.6%

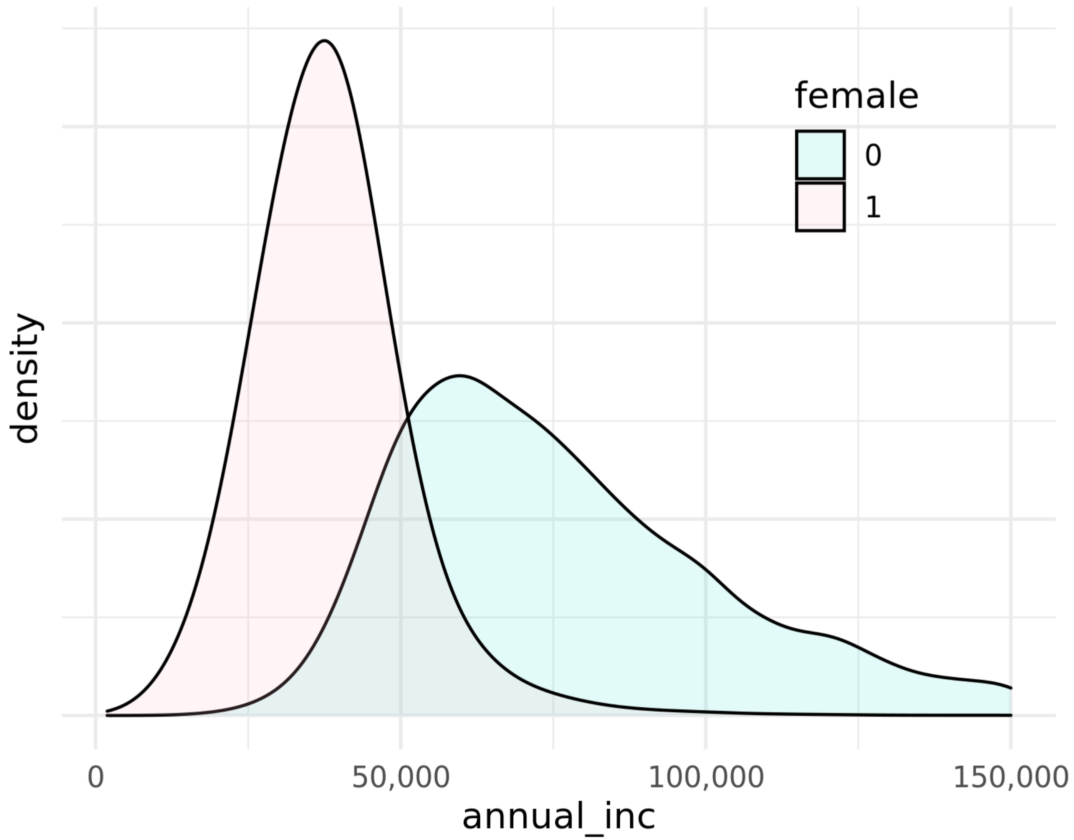
PAYCHEX®
Payroll | Benefits | HR | Insurance

Inferring “gender”

Original income distribution



Inferred “gender”



Model Building

Model A (income)

- Include income but NOT “gender”
- Use “gender” to evaluate fairness metrics

Model B (gender)

- Include “gender” but NOT income
- Use “gender” to evaluate fairness metrics

Both
- XGBoost
- 60% train, 15% test (tuning), 25% validation
  - Same splits for both
- Fairness metrics evaluated on validation data

Model Performance

Model A (income)

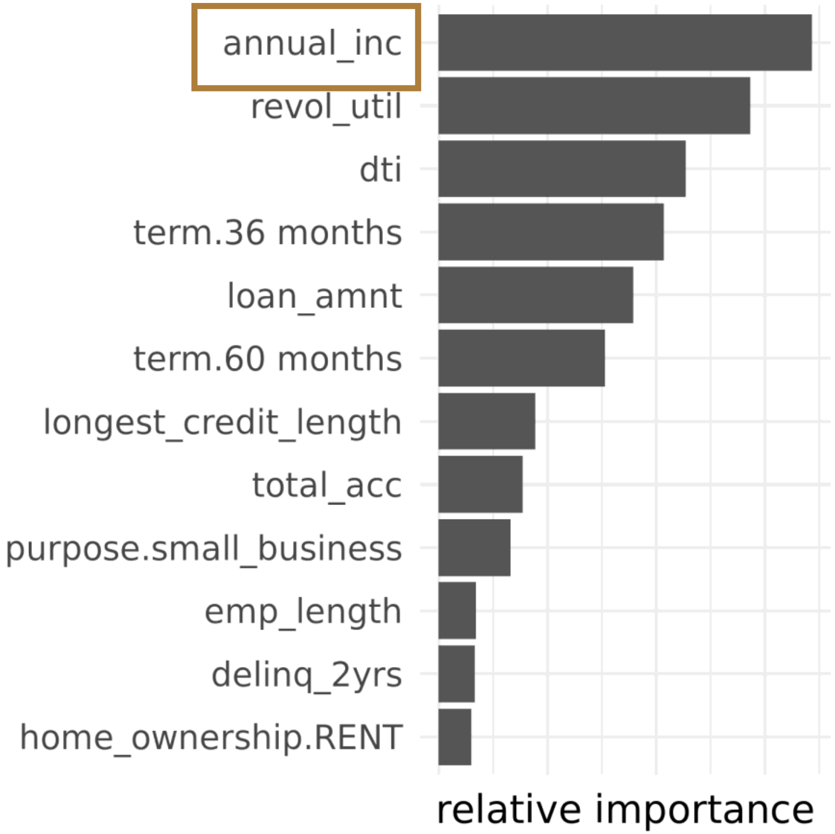
- ROC-AUC: 0.686
- PR-AUC: 0.318
- Accuracy: 66.4%

Model B (gender)

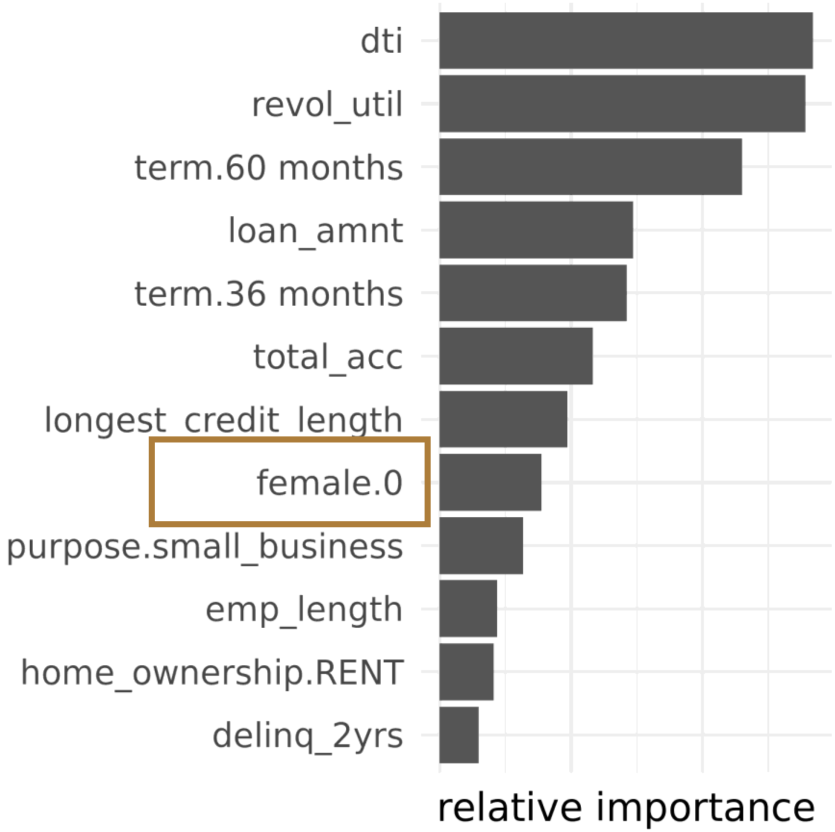
- ROC-AUC: 0.678
- PR-AUC: 0.310
- Accuracy: 65.2%

Global Importances

Model A (income)



Model B (gender)



Payroll | Benefits | HR | Insurance

Fairness Metrics

- Demographic Parity
- Calibration
- Model performance by subgroup
- Classification parity

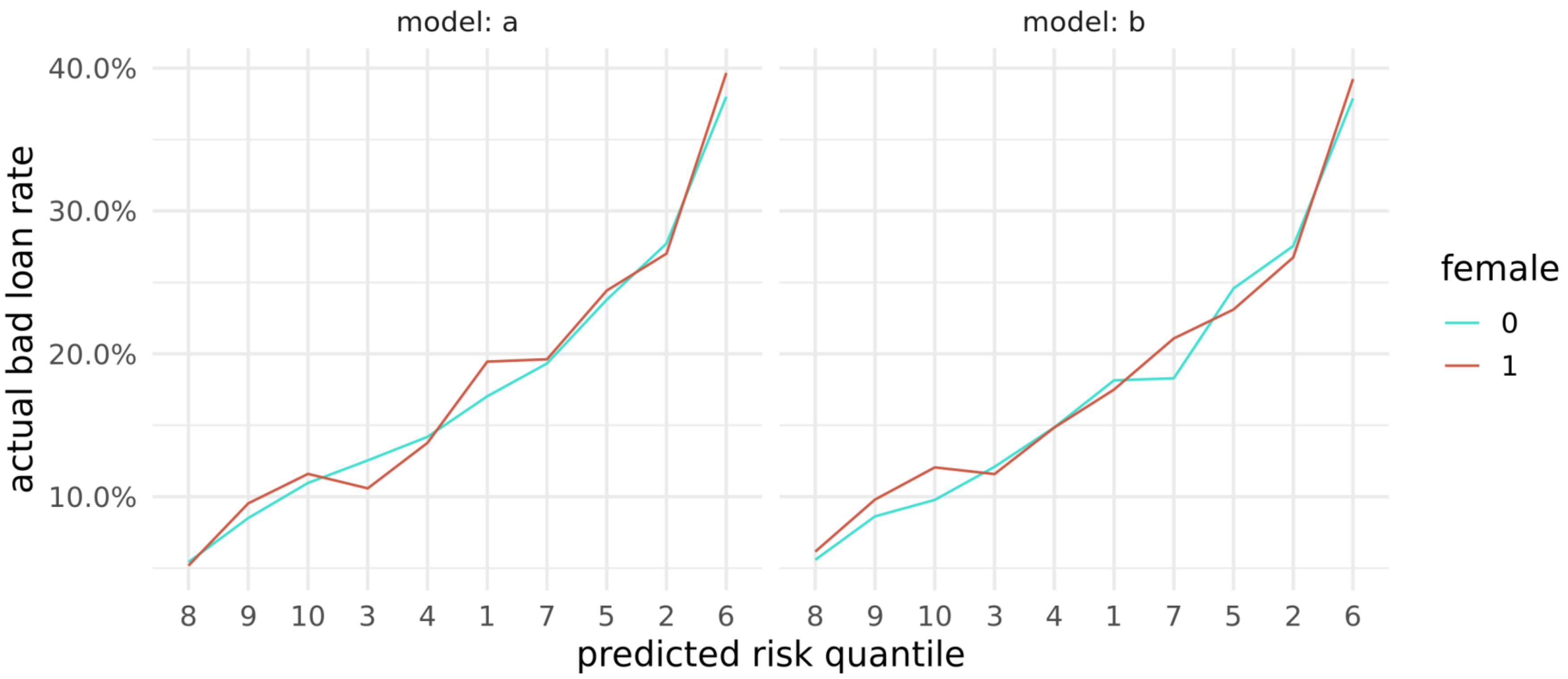
Demographic Parity

- Outcomes similar across groups?

	Model A	Model B	Actual
F	22.0%	22.1%	21.9%
M	16.7%	16.7%	16.4%

Calibration

- Similar relationship between actual and predicted risk by groups
- May examine predicted risk buckets (e.g. deciles)



Model Performance: AUC

How well does the model separate targets and non-targets, for each category?

ROC-AUC

	Model A	Model B
F	0.659	0.648
M	0.689	0.681
ALL	0.686	0.678

PR-AUC

	Model A	Model B
F	0.345	0.337
M	0.303	0.296
ALL	0.318	0.310

PAYCHEX
Payroll | Benefits | HR | Insurance

Classification Parity

- Similar confusion matrix metrics across groups
- Frequently cited
  - Accuracy
  - F1
  - False positive rate
  - Equal opportunity (non-discrimination in “desirable” outcome)
- **Impossible to satisfy both classification parity (FP or EO) and calibration when underlying base rates differ**
  - G Pleiss, M Raghavan, F Wu, J Kleinberg, KQ Weinberger. Advances in Neural Information Processing Systems, 5680-5689, 2017
  - https://arxiv.org/abs/1709.02012

Classification Parity: Accuracy and F1

How often is the model right?

Accuracy

	Model A	**Model B**
F	54.8%	52.7%
M	70.6%	69.7%
ALL	66.4%	65.2%

f1

	Model A	**Model B**
F	0.412	0.404
M	0.369	0.361
ALL	0.385	0.377

Classification Parity: False Positive Rate and "Equal Opportunity"

- FP rate: The likelihood that someone will be labelled risky when they don't actually default
- EO: The likelihood that someone will be labeled not-risky, when they don't actually default (true negative rate)

False Positives

	Model A	**Model B**
F	50.1%	53.0%
M	25.9%	26.9%
ALL	32.0%	33.5%

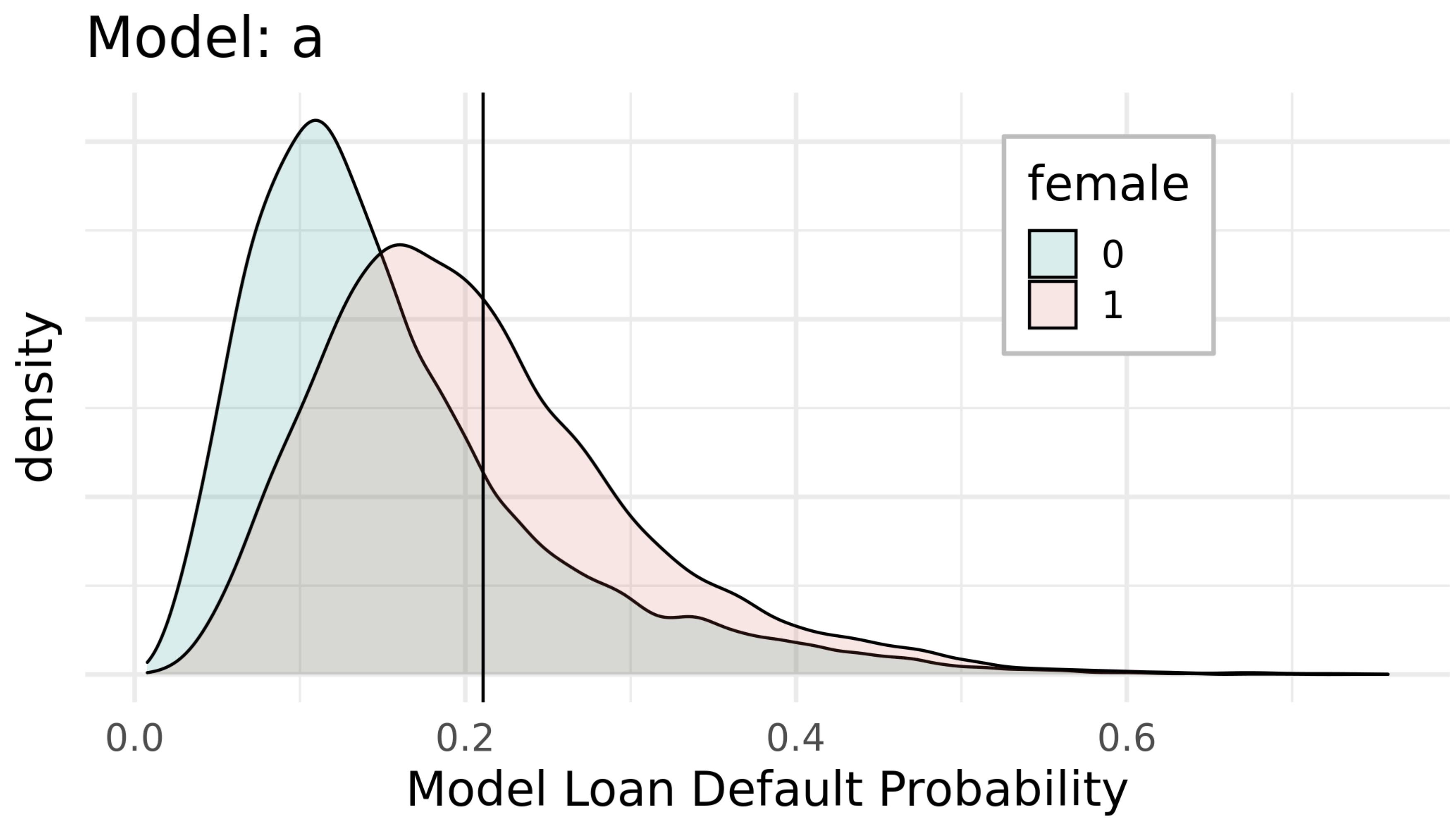
Equal Opportunity

	Model A	**Model B**
F	49.9%	47.0%
M	74.1%	73.1%
ALL	68.0%	66.5%

Females are twice as likely to be denied loans when they would actually have repaid them!

PAYCHEX
Payroll | Benefits | HR | Insurance

False Positive Rate



False positives!

G Pleiss, M Raghavan, F Wu, J Kleinberg, KQ Weinberger. Advances in Neural Information Processing Systems, 5680-5689, 2017 https://arxiv.org/abs/1709.02012

Metrics Summary

	Model A (income)	Model B (gender)
Demographic parity	FAIL (similar to actual)	FAIL (similar to actual)
Calibration	PASS	PASS
Performance (AUC, PR-AUC)	MARGINAL	MARGINAL
Accuracy	FAIL (somewhat)	FAIL (somewhat)
f1	PASS	PASS
FP rate / equal opportunity	FAIL (badly)	FAIL (badly)

So, is it fair or unfair?

Is it a stereotype or reasonable decision basis?

PAYCHEX
Payroll | Benefits | HR | Insurance

Where Are We?

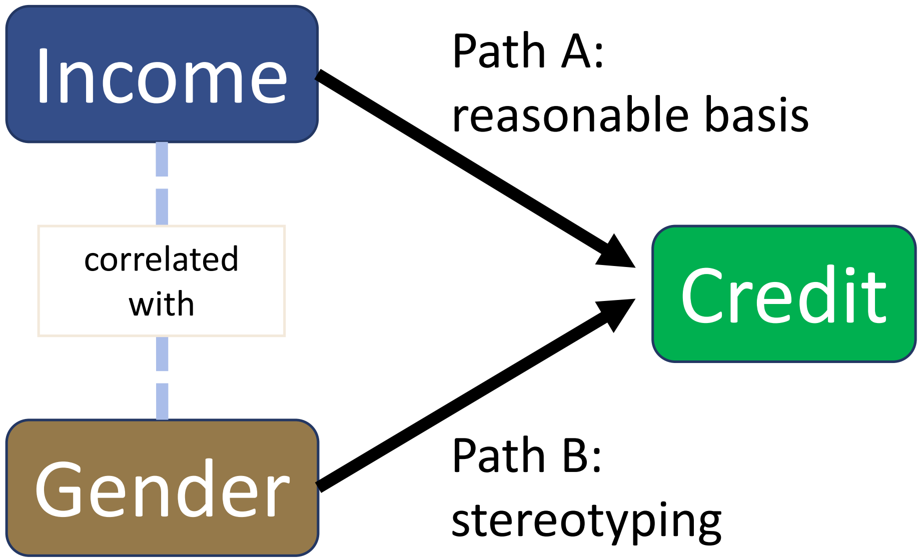
Actual Outcome Differs By Group?

Model Outcome Differs By Group?	N	Y
N	1 **Metrics Pass**	2 Calibration fail FP test fail
Y	3 Calibration fail FP test fail	4 Calibration pass FP test fail **YOU ARE HERE!**

Features

Problem

How (or can we) distinguish Path A from Path B?



A script (parts 1 & 2)

“

1. Actual and model outcomes both vary across [GROUPS]. This model [IS/IS NOT] calibrated, but fails classification parity metrics, particularly [METRICS].

2. The main features driving the group differences are [FEATURES]. We believe that the use of these features is reasonable for this problem because [REASONS].

”

Payroll | Benefits | HR | Insurance

A script (parts 1 & 2)

“

1. Actual and model outcomes both vary across genders. This model is calibrated, but fails classification parity metrics, particularly equal outcomes and false positive rate parity.

2. The main features driving the group differences are [FEATURES]. We believe that the use of these features is reasonable for this problem because [REASONS].

”

PAYCHEX®
Payroll | Benefits | HR | Insurance

Finding discrepancy-associated features

> The main features driving the group differences are [FEATURES]. We believe that the use of these features is reasonable for this problem because [REASONS].

Which features are “responsible” for the difference in predictions for men vs. women?

- Shapley explanations (tailored)
- Inductive reasoning

Shapley

"…The feature values enter a room in random order….. The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them."

Christoph Molnar

"A Guide for Making Black Box Models Explainable"

https://christophm.github.io/interpretable-ml-book/shapley.html

Problem : A sales team of three receives a commission of $90k. How do we distribute that $90k among employees A, B and C ?

Solution: Examine all sales, and calculate the average marginal value of including vs. excluding the employee.

Shapley

Problem : A sales team of three receives a commission of $90k. How do we distribute that $90k among employees A, B and C ?

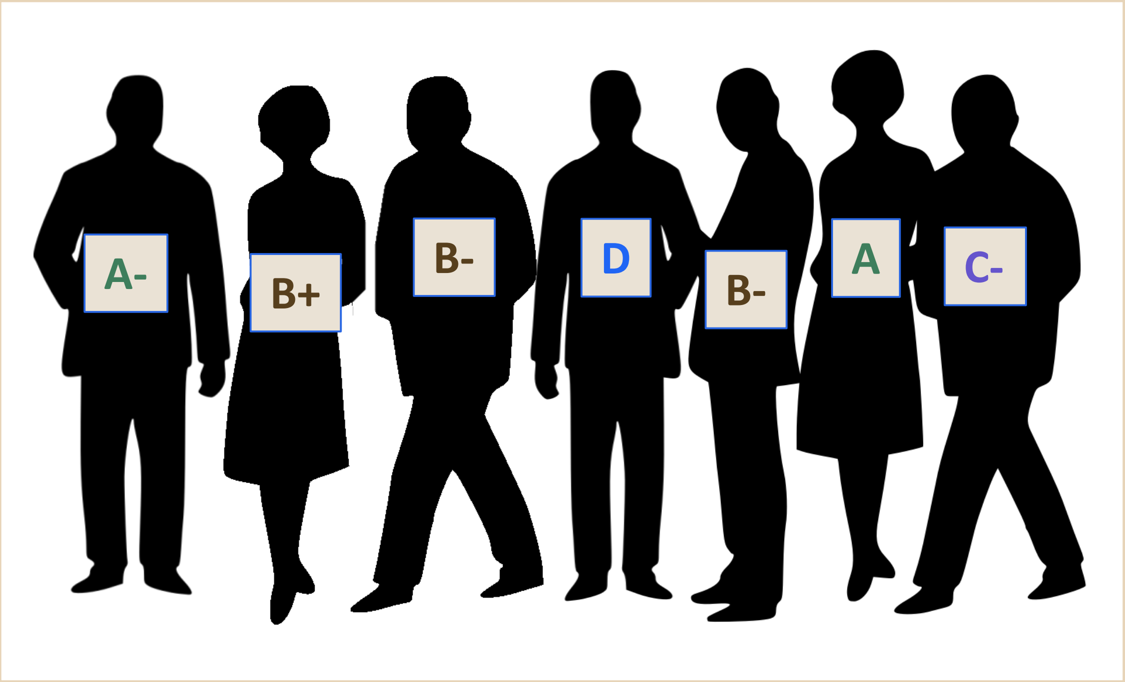
$$
V(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A,B\} \\ 85, & \text{if } c = \{A,C\} \\ 72, & \text{if } c = \{B,C\} \\ 90, & \text{if } c = \{A,B,C\} \end{cases}
$$

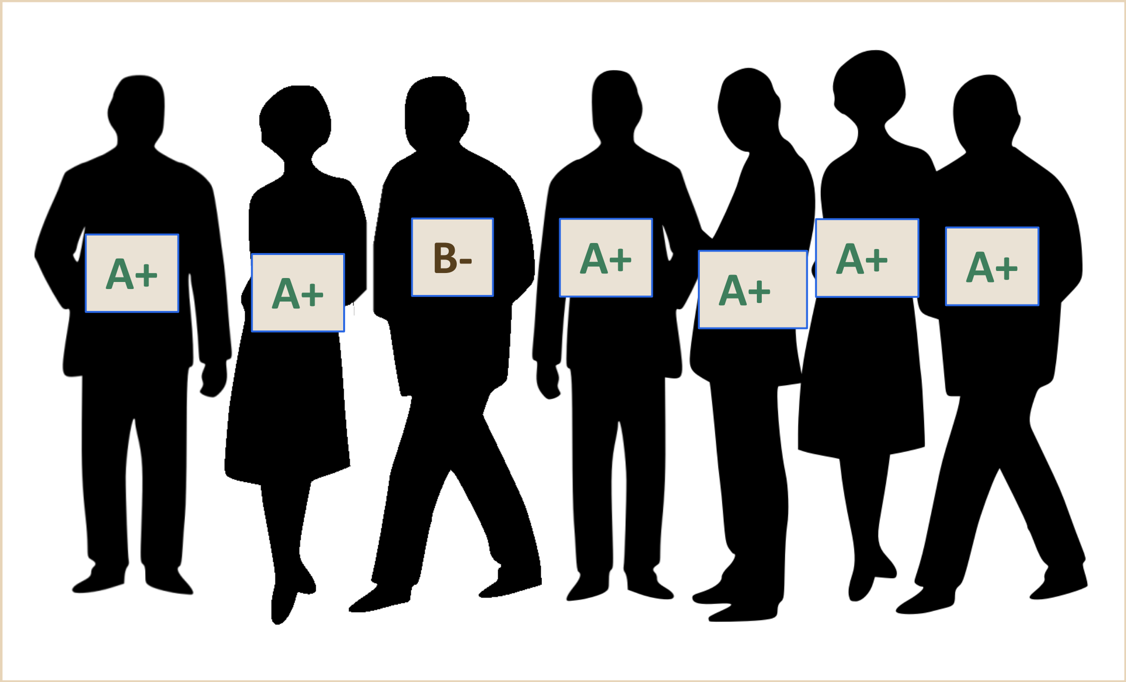
Set	Marginal Contribution
(A,B,C)	( )
(A,C,B)	(80 5 5)
(B,A,C)	(24 56 10)
(B,C,A)	(18 56 16)
(C,A,B)	(15 5 70)
(C,B,A)	(18 2 70)
Phi (Shapley Value)	**(39.2 20.7 30.2)**

- This empty room really is empty… our baseline is zero!
- For ML models phi values are relative to a baseline
- **We can choose our baseline and make explanations contrastive!**

PAYCHEX®
Payroll | Benefits | HR | Insurance

Which room am I walking into?





PAYCHEX
Payroll | Benefits | HR | Insurance

Allocating Disparities using Shapley

If we use data from males to define our "empty room", we can find excess probability for females relative to this population.

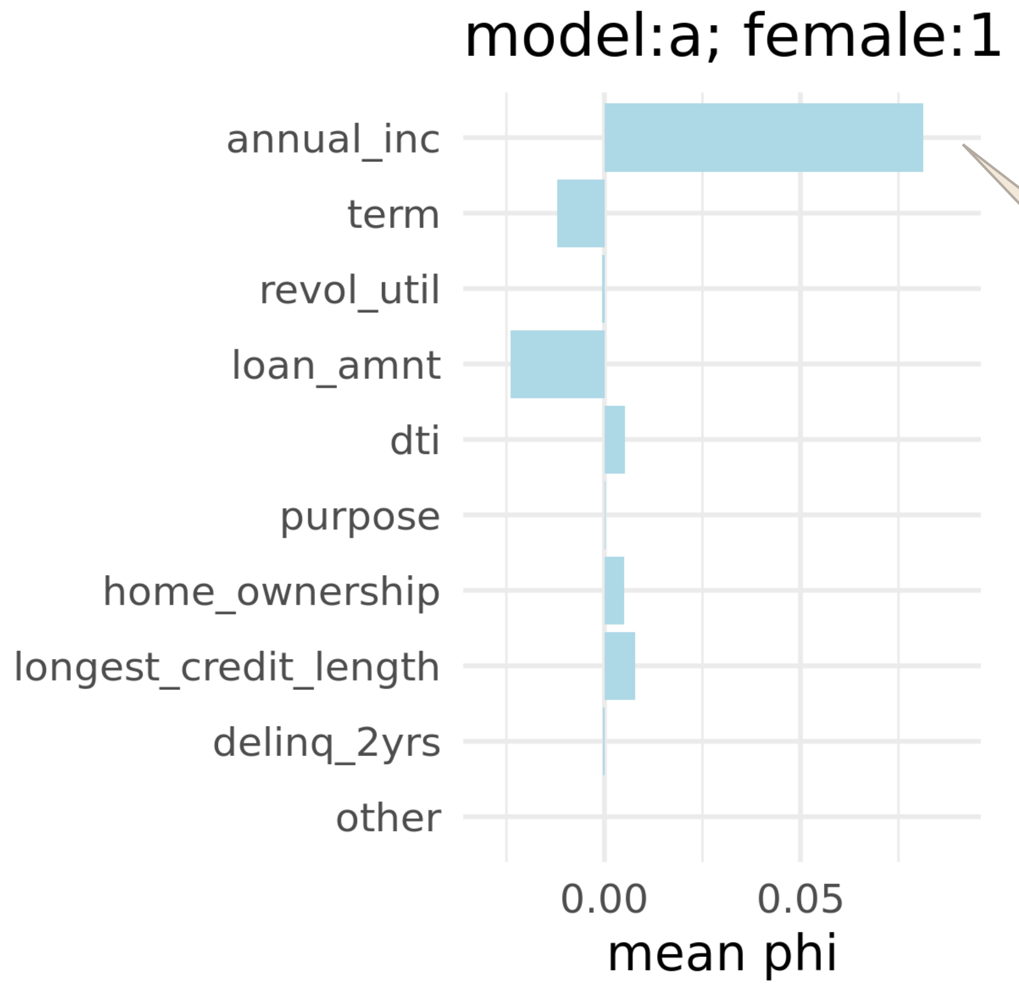
If we aggregate individual probabilities for females, we can measure the effects of each feature on the female population relative to the male

- Individual probabilities can be summed to find the group rates. Therefore, individual Shapley values can be summed to find the contribution of a feature to a group's rate!
- Use training data

- See: Explaining Measures of Fairness by Scott Lundberg
  - https://towardsdatascience.com/explaining-measures-of-fairness-f0e419d4e0d7
  - Different baseline data sets; specific to stereotyping

PAYCHEX®
Payroll | Benefits | HR | Insurance

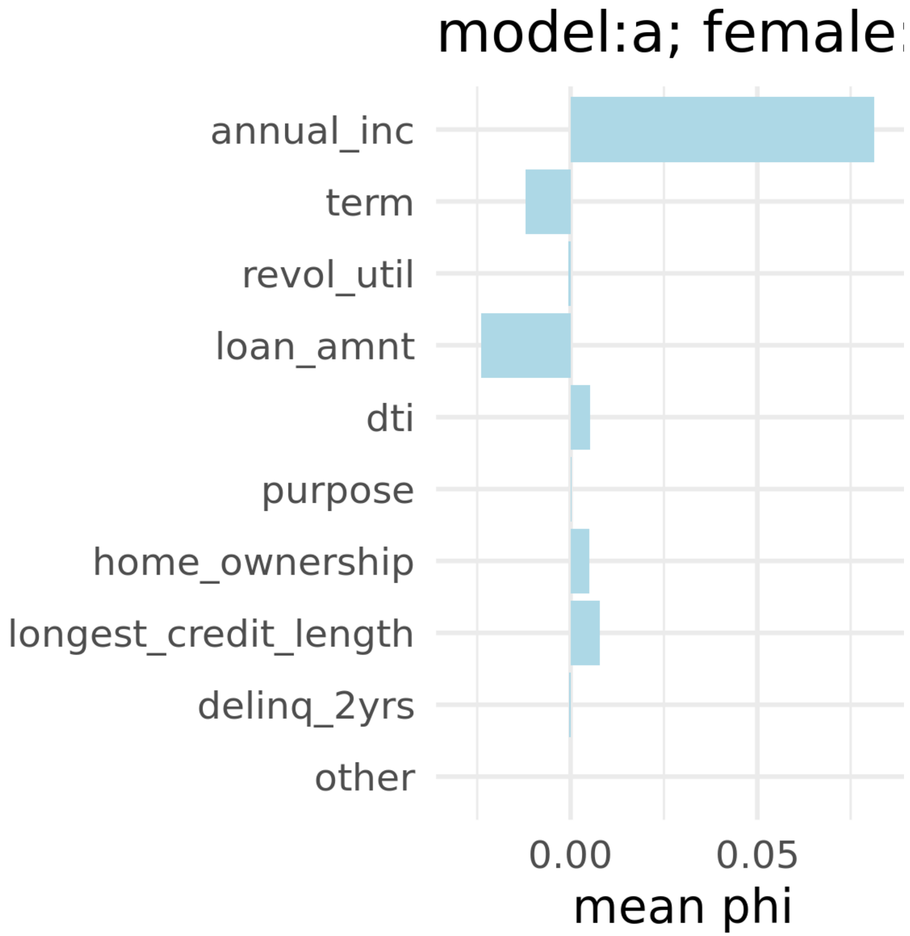
Model A

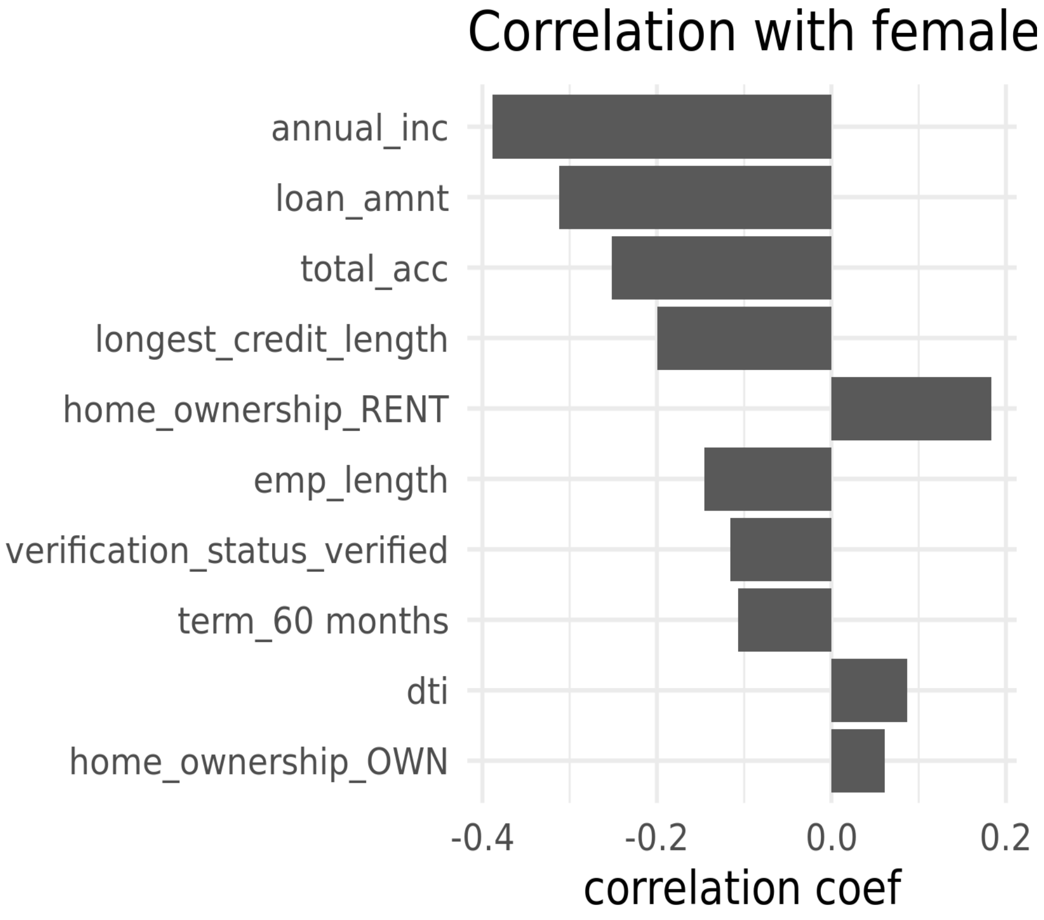


Mean p1 for females (sample):	23.1%
Mean p1 for male “foil” (sample):	16.8%
Delta:	**6.3%**
Sum of Shapley means:	**6.3%**

Positive values are most important to explain increase in defaults for females

Model A: Shapley vs. Correlations



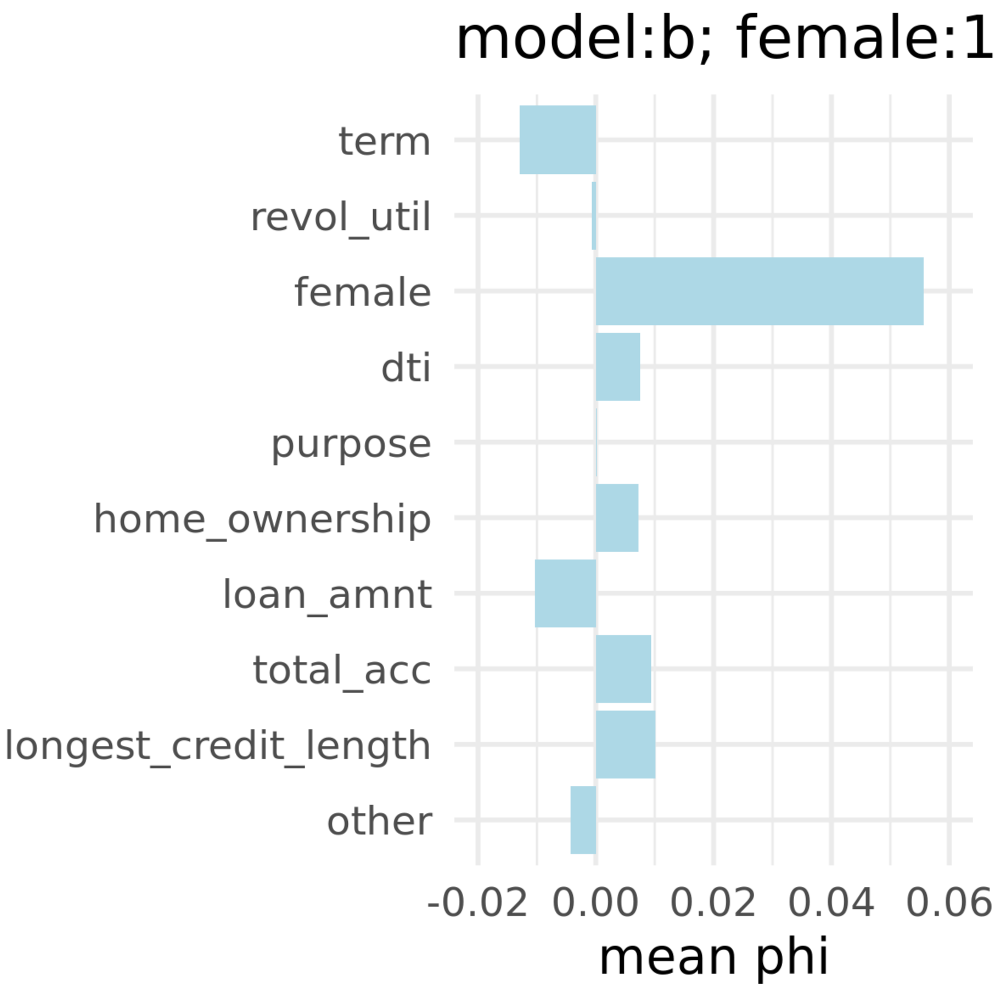


Shapley isolates feature combinations to the model!

Specifically, features contributing to excess probability relative to the reference!

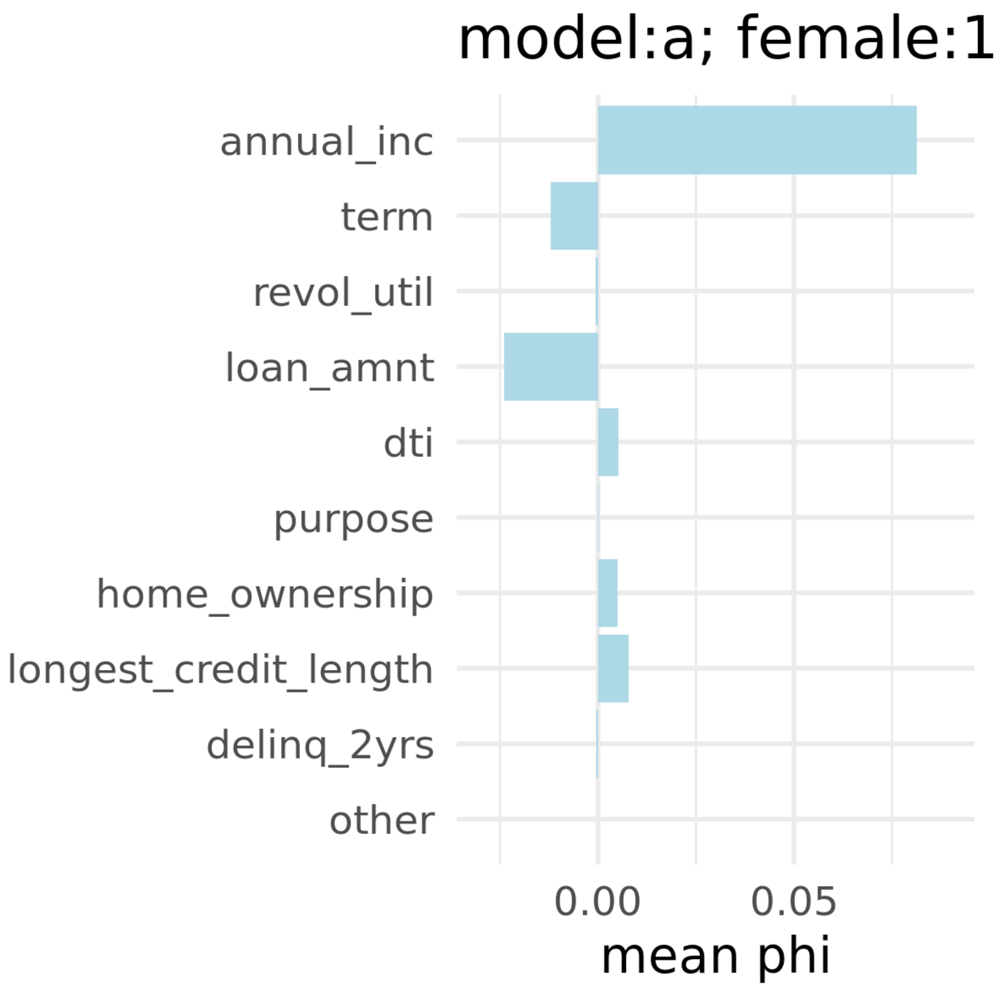
PAYCHEX
Payroll | Benefits | HR | Insurance

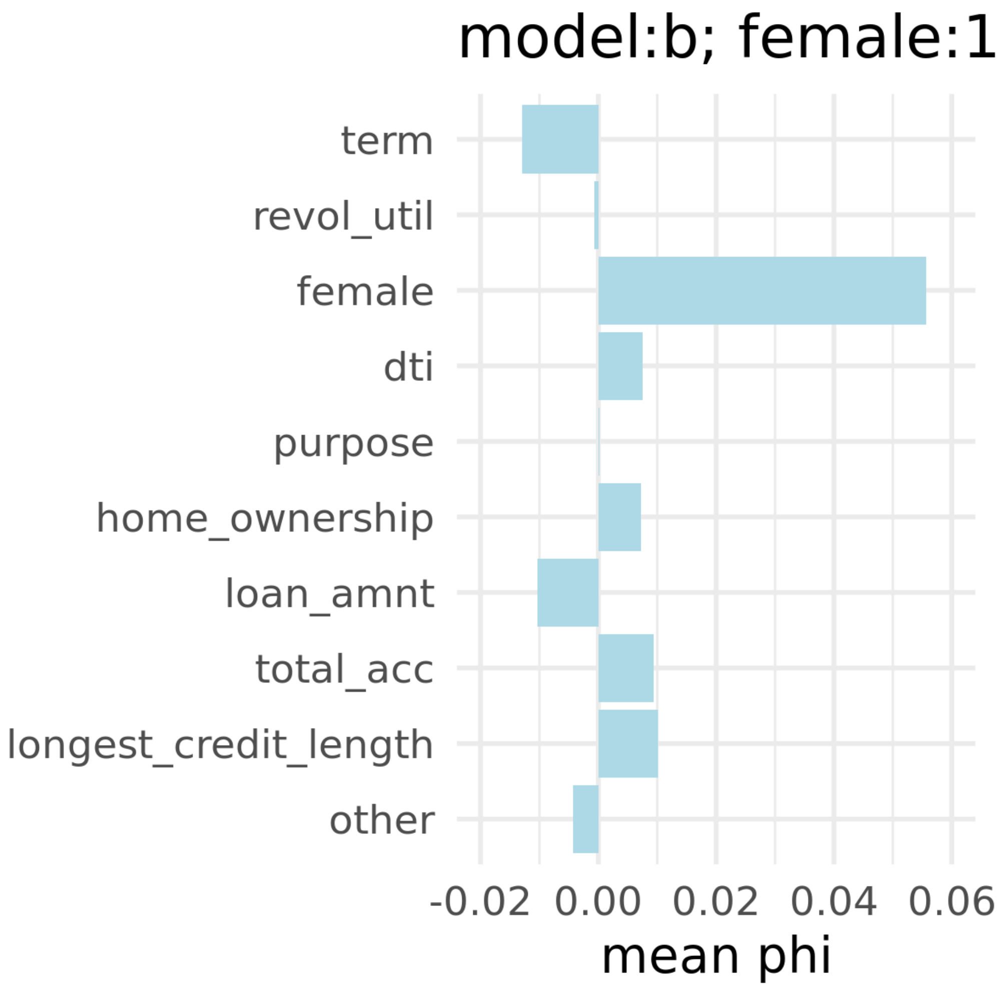
Model B



Mean p1 for females (sample):	22.8%
Mean p1 for male “foil” (sample):	16.7%
Delta:	**6.1%**
Sum of Shapley means:	**6.2%**

Now what?


model:a; female:1
annual_inc
term
revol_util
loan_amnt
dti
purpose
home_ownership
longest_credit_length
delinq_2yrs
other
0.00
0.05
mean phi


model:b; female:1
term
revol_util
female
dti
purpose
home_ownership
loan_amnt
total_acc
longest_credit_length
other
-0.02
0.00
0.02
0.04
0.06
mean phi

Why this feature?

Possible Reasons

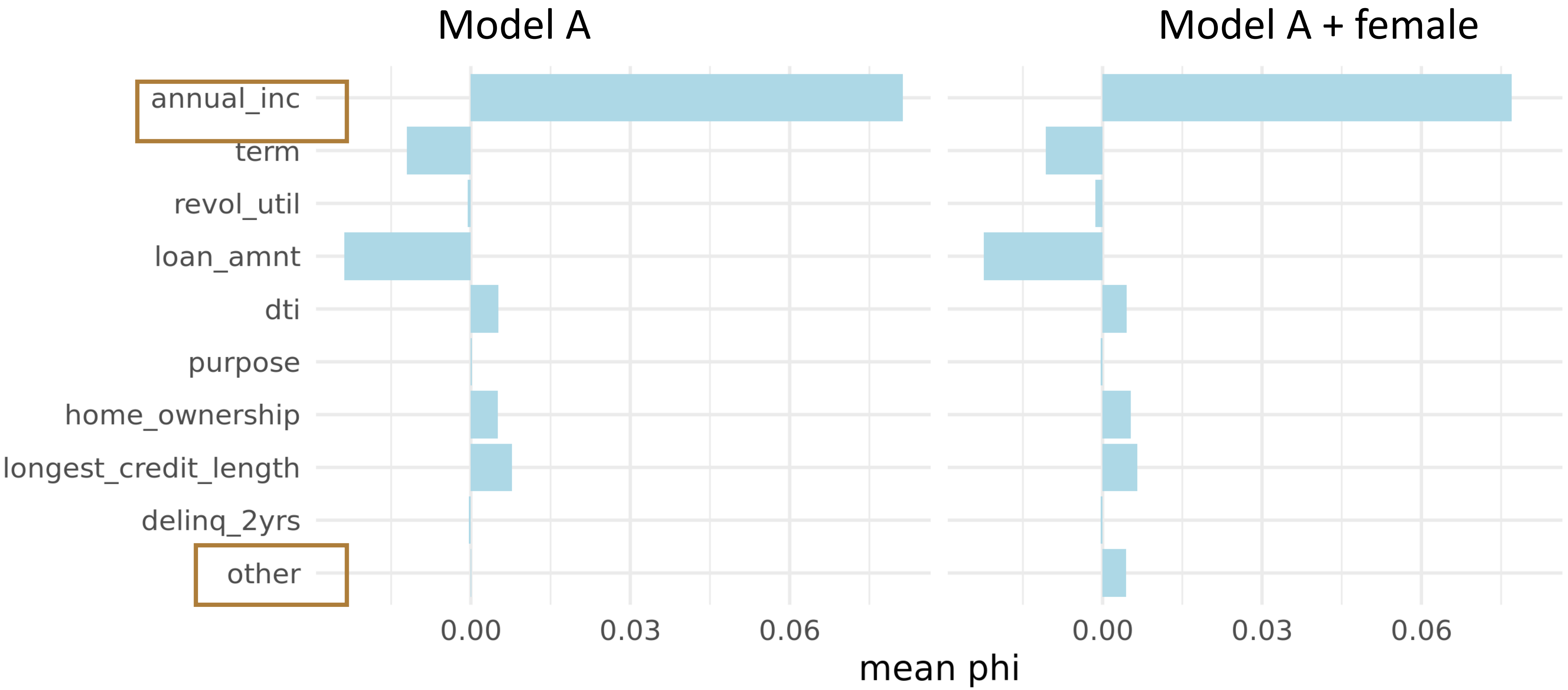
1. The feature is relevant and independently associated with the outcome
2. The feature is a proxy for some missing variable, which is also correlated with group membership (stereotyping)
3. Label bias
4. Feature bias

Tests

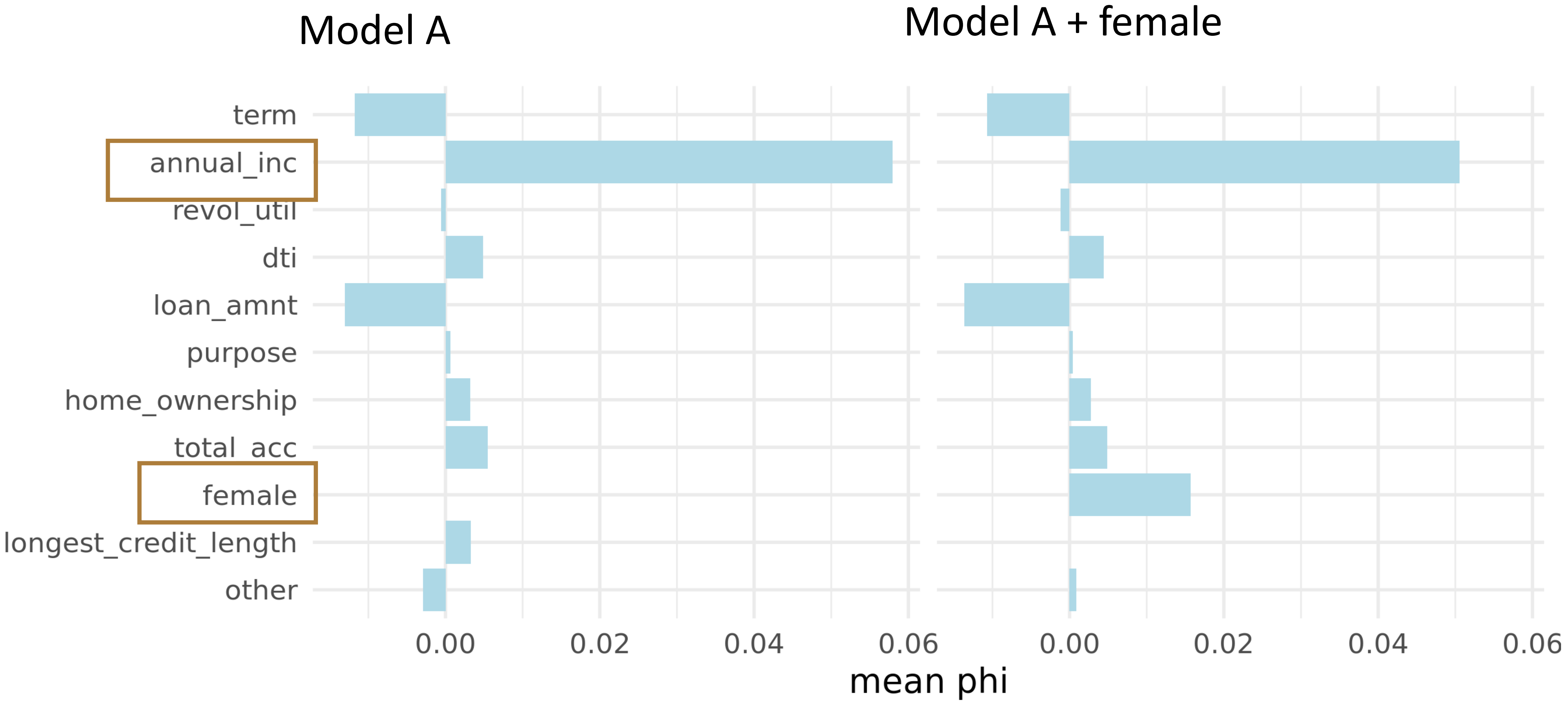
1. Create similar with sensitive feature
2. Model with sensitive feature, and toggle feature value
3. Use the model output as a predictor in a new model with and without sensitive feature

PAYCHEX
Payroll | Benefits | HR | Insurance

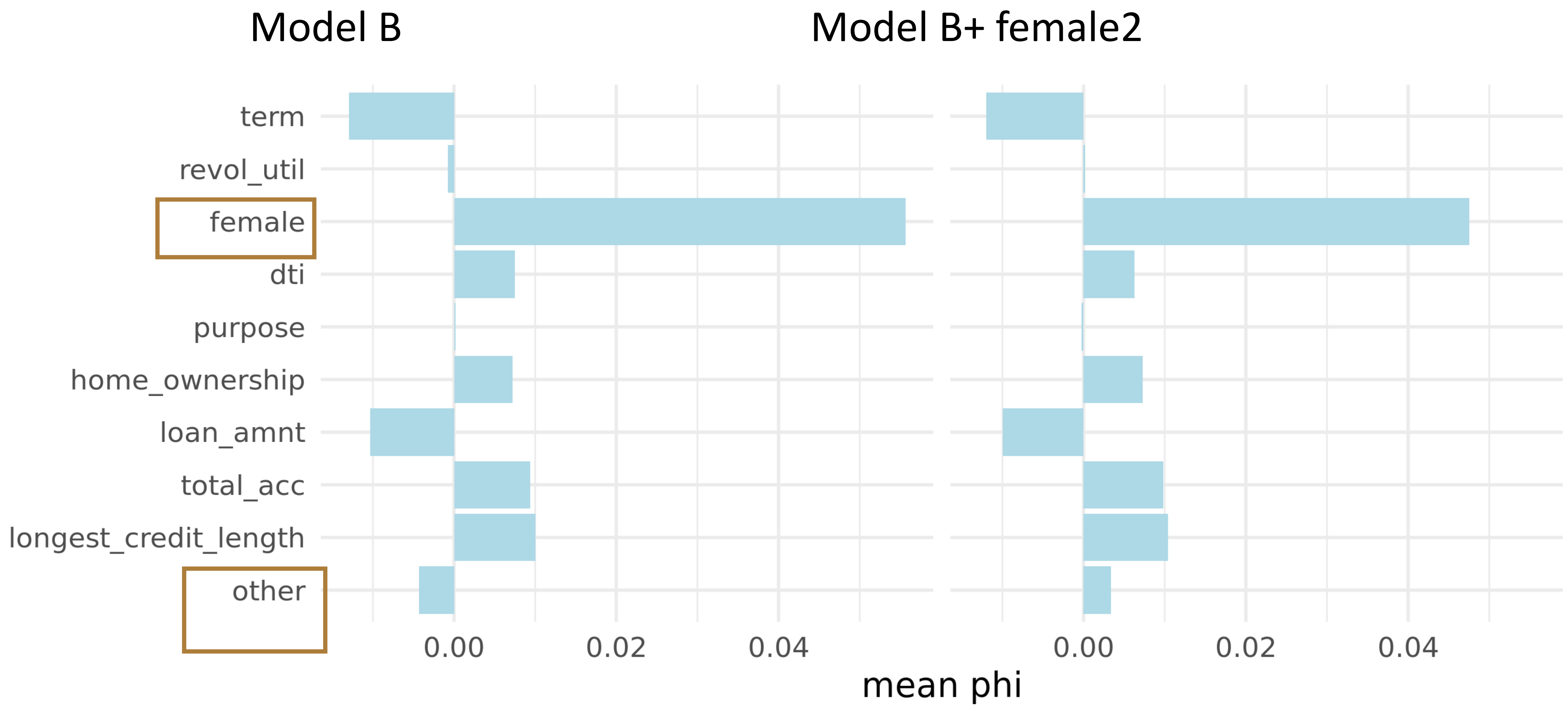
Model A + female



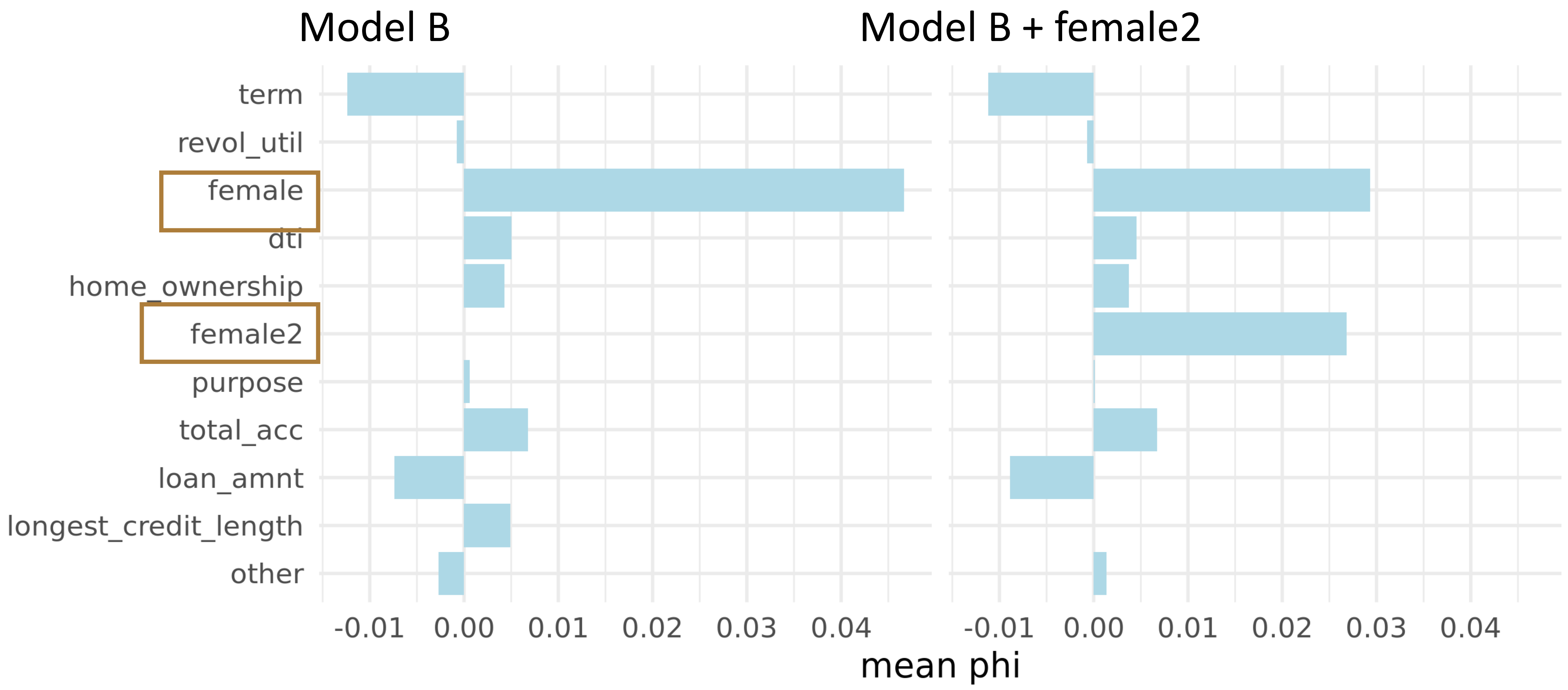
Model A + female - Random Forest



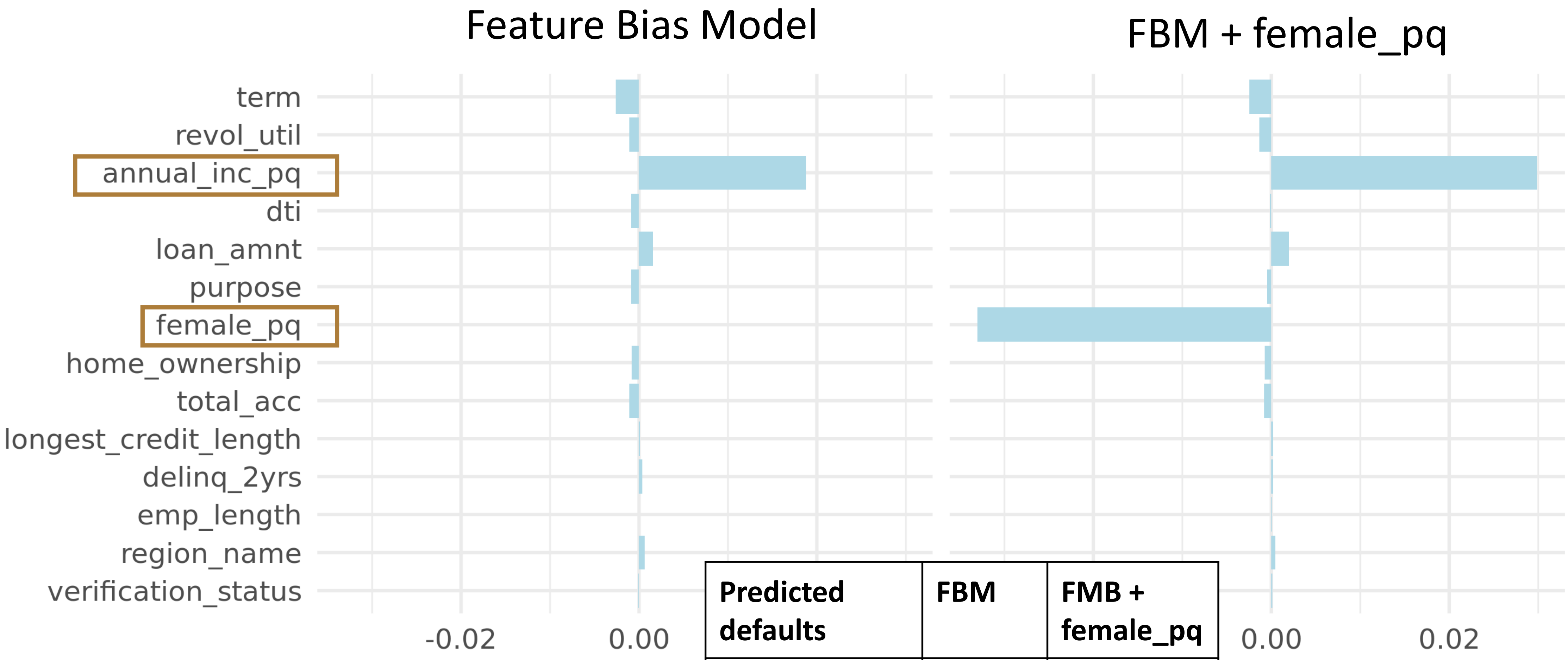
Model B + female2



Model B + female2 - Random Forest

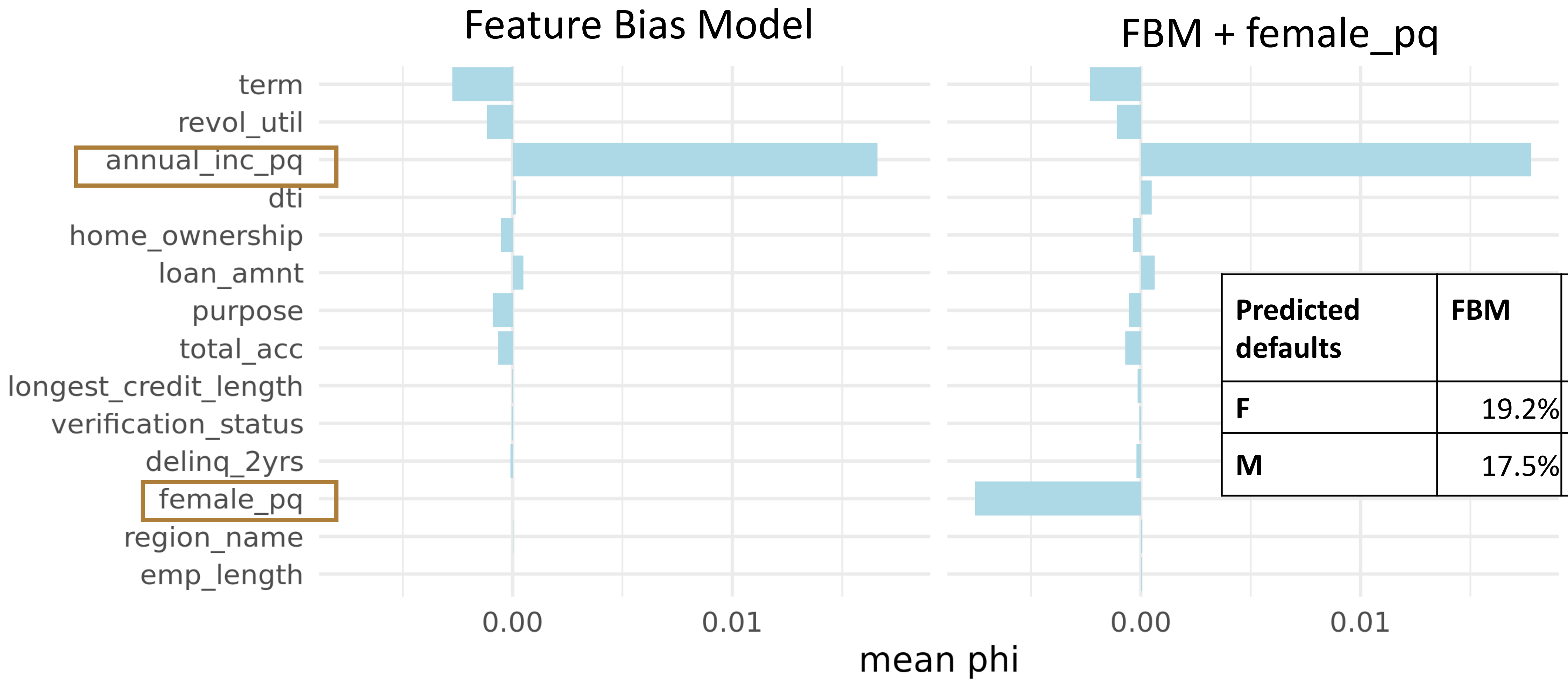


Feature bias



Predicted defaults	**FBM**	**FMB + female_pq**
F	19.2%	18.0%
M	17.4%	18.2%

Feature bias - Random Forest



Predicted defaults	FBM	FMB + female_pq
F	19.2%	18.9%
M	17.5%	17.7%

A script (parts 1 & 2)

“

1. Actual and model outcomes both vary across genders. This model is calibrated, but fails classification parity metrics, particularly equal outcomes and false positive rate parity.

2. The main features driving the group differences are [FEATURES]. We believe that the use of these features is reasonable for this problem because [REASONS].

”

PAYCHEX®
Payroll | Benefits | HR | Insurance

A script (part 2)

Model A

2. The main feature driving the group differences is **annual income**. We believe that the use of this feature is reasonable for this problem because **it is directly related to a person's ability to repay a loan, and tests show stereotyping is unlikely**

Model B

2. The main feature driving the group differences is **female status**. We believe that the use of this feature is reasonable for this problem because **????**

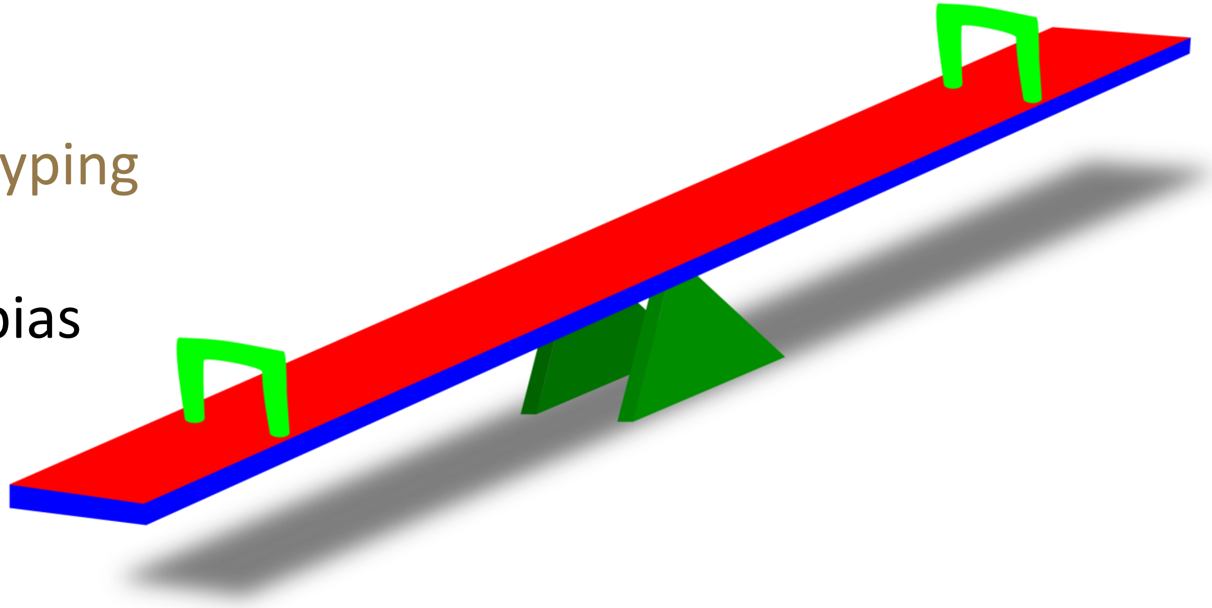
PAYCHEX®
Payroll | Benefits | HR | Insurance

Sensitive Features

Sensitive features

Leave them out!

- Reduce stereotyping risk
- Mitigate label bias



Put them in!

- Mitigate feature bias
- Increase overall accuracy

A rising tide lifts all boats

Scenario 1

	FP Rate
F	28%
M	29%
ALL	**29%**

Scenario 2

	FP Rate
F	5%
M	1%
ALL	**3%**

Low-risk women are 5 times more likely to be unfairly denied!!

But they much are better off overall

Payroll | Benefits | HR | Insurance

Model B – should we include female status?

	ROC-AUC	Overall FP rate	Male FP rate	Female FP rate
Model B	0.678	33%	27%	53%
Model B - female	0.674	30%	29%	34%

Probably not…?

	Overall FN rate	Male FN rate	Female FN rate
Model B	41%	48%	27%
Model B - female	45%	45%	45%

More equal, but worse for females?

PAYCHEX
Payroll | Benefits | HR | Insurance

Focus & Burden

Spelman College®
A Choice to Change the World

"Digitizing Discrimination"

Brandeis Marshall, Ph.D.

@ University of Virginia

Oct 6, 2020 11:00 AM EST

UNIVERSITY of VIRGINIA

PRINCETON UNIVERSITY

"Algorithmic Fairness and Decision Landscapes"

Annette Zimmermann, Ph.D.

Algorithmic Ethics: Perspectives from Philosophy and Computer Science Workshop

@ University of Rochester

May 1, 2020 11:00 AM EST

UNIVERSITY of ROCHESTER

PAYCHEX®
Payroll | Benefits | HR | Insurance

Focus & Burden

	FP Rates		FN Rates	
	M	F	M	F
Model B	27%	53%	48%	27%
Model B - female	29%	34%	45%	45%

Goal	Best model	Who carries the burden?
Make sure as many "deserving" women as possible get loans	Model B - female	• Women who are given loans they can't afford (lower income women)
Prevent defaults among vulnerable women	Model B	• Women who would have paid their loans but are denied (higher income women)

PAYCHEX® Payroll \| Benefits \| HR \| Insurance

A script (parts 3 & 4)

“

3. Additional features that might improve the outcome and reduce unfairness include [FEATURES]. It is reasonable in our situation to proceed without these because [REASONS].

4. We believe that inclusion of sensitive features [IS/IS NOT] beneficial for this model because [REASONS].

”

What else?

> 3. Additional features that might improve the outcome and reduce unfairness include [FEATURES]. It is reasonable in our situation to proceed without these because [REASONS].

Omitting features with high predictive value and/or causal relationships with the outcome risks:

- No-win tradeoffs
- Stereotyping (direct or by proxy)

Sensitive features

> “We believe that inclusion of sensitive features [IS/IS NOT] beneficial for this model because [REASONS].”

- There are **tradeoffs** to consider in inclusion decisions
  - Including mitigates predictor bias and may improve overall accuracy
  - Removal reduces stereotyping risk and may mitigate label bias in specific circumstances

PAYCHEX®
Payroll | Benefits | HR | Insurance

Summary

- Fairness metrics allow you to assess overall performance, calibration, and potential risks in your models
- Fairness metrics **do not** distinguish between stereotypes and reasonable decision bases
- Shapley values enable discovery of features driving differences between groups, and some assessments of type of bias
- Sensitive features may improve or reduce fairness

A script

“

1. Actual and model outcomes both vary across [GROUPS]. This model [IS/IS NOT] calibrated, but fails classification parity metrics, particularly [METRICS].

2. The main features driving the group differences are [FEATURES]. We believe that the use of these features is reasonable for this problem because [REASONS].

3. Additional features that might improve the outcome and reduce unfairness include [FEATURES]. It is reasonable in our situation to proceed without these because [REASONS].

4. We believe that inclusion of sensitive features [IS/IS NOT] beneficial for this model because [REASONS].

”

PAYCHEX®
Payroll | Benefits | HR | Insurance

“The fundamental difference is the amount of thoughtfulness built in”

“Forget the robots! Here’s how AI will get you”
Cassie Kozyrkov
https://towardsdatascience.com/forget-the-robots-heres-how-ai-will-get-you-b674c28d6a34

“The AI safety mindset: 12 rules for a safer AI future”
Cassie Kozyrkov
https://www.youtube.com/watch?v=EjBXZrQ7fTs

Comments? Questions?