ORCAA

Explainable Fairness: A Framework

April 26, 2022

Everyone agrees AI should be “fair.”

But what’s “fair”?

What’s Happening Now

Government is starting to demand “unbiased” AI.

- New laws: NYC, DC, CO
- Regulated industries first
- Key definitions left to rulemaking

Companies are improvising since they don’t know how to prepare.

- Voluntary “exhibition” audits
- No-action letters
- One-off approaches e.g., Twitter, Airbnb

What’s missing?

Clear definitions, rules, and standards.

1. Explainable Fairness framework

There is no universal definition of “fairness,” even in the context of existing antidiscrimination laws.

We are proposing what we hope is a reliable way to

- Come up with a precise definition for a given scenario,
- Test an AI system against that definition, and
- Produce evidence of fairness that a nontechnical person can understand.

We call this way **Explainable Fairness**.

How To Explain Fairness

The basic idea is: no ethnic group (or gender group, age group, etc.) should fare better than others on average, unless there is a valid explanation in the form of an attribute or set of attributes that “accounts for” the difference.

The process is:

1. Define an outcome of interest: who is eligible for a loan, or insurance, for example.
2. Infer protected class status of individuals, like race or gender, knowing no method is perfect.
3. Measure the outcomes for each of the protected classes.
4. If there is a difference, could other (non-protected) characteristics “account for” some of it? If so, make an argument for each.
5. Re-measure outcomes and differences after accounting for agreed control(s).

Explainable Fairness Solves a Regulator’s Problems

ORCAA

Balancing Framework

1. Regulator: There's a difference in outcome
2. Industry: Oh but there's such a good reason for it, i.e. a benefit
3. Regulator: Even accounting for those beneficial factors, we still see a difference, we need to balance.

Explainable Fairness helps create a positive feedback loop

If a regulator does the same balancing test for the entire industry, they get to set and, over time, improve standards.

Imagine a “Consumer Reports” View for Regulators

ORCAA

What’s Happening Now

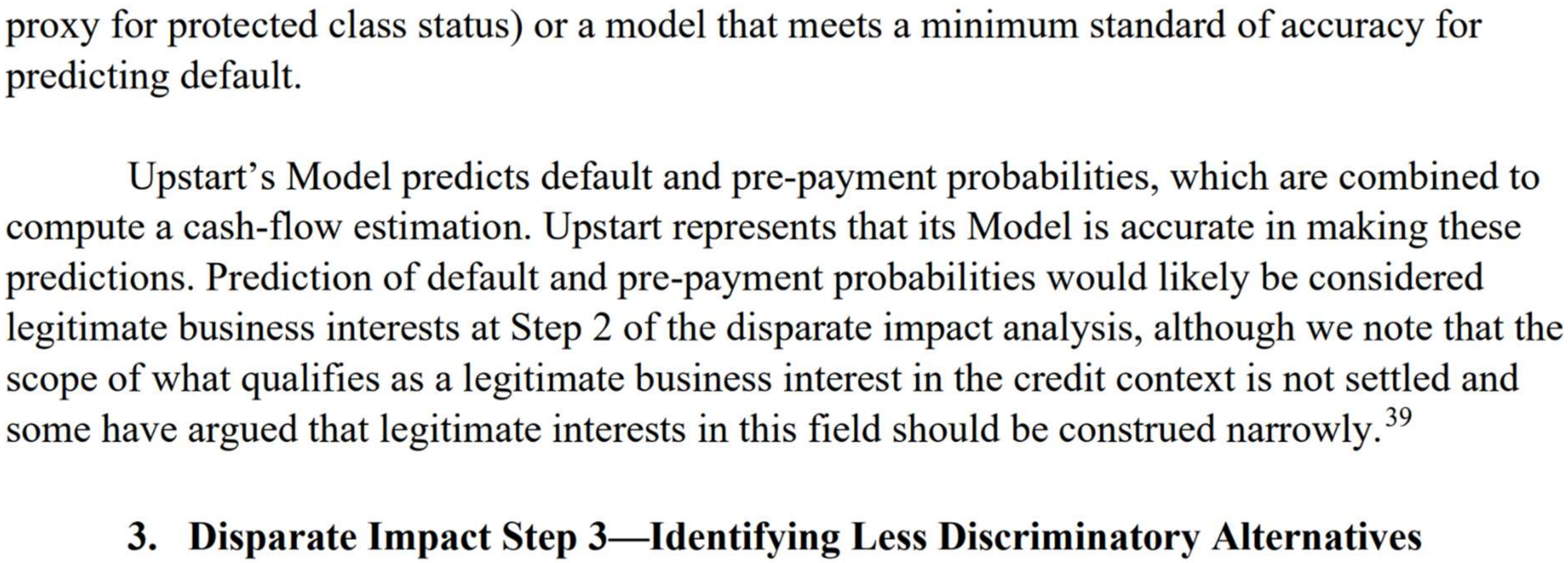
That’s where Pilot, or other auditing platforms, come in.

PILOT

They do the fairness testing to show regulators the “stragglers”

Here’s how not to do the second step…

Relman Colfax's report on Upstart



proxy for protected class status) or a model that meets a minimum standard of accuracy for predicting default.

Upstart's Model predicts default and pre-payment probabilities, which are combined to compute a cash-flow estimation. Upstart represents that its Model is accurate in making these predictions. Prediction of default and pre-payment probabilities would likely be considered legitimate business interests at Step 2 of the disparate impact analysis, although we note that the scope of what qualifies as a legitimate business interest in the credit context is not settled and some have argued that legitimate interests in this field should be construed narrowly.[39]

3. Disparate Impact Step 3—Identifying Less Discriminatory Alternatives

Use Case: Student Lending

Step 1: Outcomes of interest include loan offer, APR, term, consequences of late payment or default. Choose one.

Step 2: Infer protected class status and categorization buckets.

Step 3: Measure outcomes for various classes.

Step 4: If we see “significant”* differences, consider “accounting factors”** such as: FICO, salary, type of major, rank of college.

* what does significant mean here? ** when is a factor considered legitimate?

Use Case: Disability Insurance

Step 1: Outcomes of interest include: claim approval, length of initial claim, number of total weeks covered, number of total extensions.

Step 2: Infer protected class status and categorization buckets.

Step 3: Measure outcomes for various statuses.

Step 4: If we see "significant"* differences, consider "accounting factors"** such as: type of injury, maternity claim, age, comorbidity, type of work.

*what does significant mean? ** when is it considered legitimate?

Use case: auto insurance (1 of 5)

Step 1: Outcome of interest
LR = Loss Ratio = Losses / Premium

Note that LR is a helpful choice because losses “cancel out”:
Premium = Losses + Expenses + Profit →
LR = Losses / Premium
= (Premium - Expenses - Profit) / Premium
= 1 - (Expenses / Premium) - (Profit / Premium)
= 1 - Expense ratio - Profit ratio

…leaving just the expense ratio and profit ratio. Expense ratio could legitimately vary by race or other protected class (this is the argument insurers will have to make), but profit ratio probably can’t.

Use case: auto insurance (2 of 5)

Now we have to request data for testing. Scope:

- A benchmark policy or set of policies (e.g. certain coverages, limits)
- Policies with loss data (i.e., most/all claims have been settled)
- One row per policy
- Fields: name, address, premium, losses
    - Later in this example we will also want Bluebook value of the car, and customer’s CBIS score.

Step 2:
Infer race using BIFSG
Then create a binary variable:
Race = {1 if White or API; 0 otherwise}

Use case: auto insurance (3 of 5)

Step 3: Primary analysis
Linear regression.
LR = c + $\boldsymbol{\beta}_1$*Race + $\boldsymbol{\varepsilon}$

If $\boldsymbol{\beta}_1$ is either very small or non statistically significant, you’re done.

Otherwise, proceed to “Step 4: Legitimate factors”

Use case: auto insurance (4 of 5)

Step 4: Legitimate factors

1. [Certain factors may be allowed without a balancing test]
2. In general, factors must pass a balancing test to be considered

   Suppose insurer proposes CBIS as a legitimate factor. The balancing test is a ratio:

   Ratio = [how well does CBIS predict LR] / [how much does CBIS proxy for race]

   - [how well does CBIS predict LR]: linear regression
     - LR = c + $\boldsymbol{\beta}_2$*CBIS + $\boldsymbol{\varepsilon}$. The r^2 of this regression is the numerator of the ratio.
   - [how much does CBIS proxy for race]: linear regression
     - CBIS = c + $\boldsymbol{\beta}_3$*Race + $\boldsymbol{\varepsilon}$. The r^2 of this regression is the denominator of the ratio.

     Criteria for a given factor to be legitimate
   - The ratio is at least [x]
   - Denominator is at most [y]

   Additional constraint: All legitimate factors together cannot predict race beyond [z]

Slide 18

2 Insurers will probably insist on factors like value of the car being insured (since there are some fixed costs to underwriting -- e.g. there might be a minimum commission to the agent no matter how cheap the policy, or a fixed cost to look up the repair history of a VIN; these would lead to higher expense ratios for cheap cars). I think we have to allow these, even if they are highly correlated with race. What if we try a definition like “there is a clear causal relationship between this variable and the expense ratio”. But we need to figure this out.

Jacob Appel, 5/31/2022

1 We could consider other goodness-of-fit measures besides r2.

Jacob Appel, 5/31/2022

Use case: auto insurance (5 of 5)

Step 5: Secondary analysis, including legitimate factors
Continuing the example, suppose insurer successfully got 2 legitimate factors:

- Value = Bluebook value of the car
- FICO

Now we do one more regression:
LR = c + $\boldsymbol{\beta}_1$*Race + $\boldsymbol{\beta}_2$*Value + $\boldsymbol{\beta}_3$*FICO + $\boldsymbol{\varepsilon}$

If $\boldsymbol{\beta}_1$ is either very small or non statistically significant, you're done.
Otherwise, you have an unexplained disparity in outcomes.

ORCAA

ORCAA is an algorithmic auditing consultancy.

We help clients identify and manage risks related to the use of AI/ML systems – especially risks related to fairness, bias, and discrimination.

Our DNA is found in CEO Cathy O'Neil's 2016 book *Weapons of Math Destruction*.