ASPEN TECH POLICY HUB

POLICY

Fair Algorithmic Housing Loans

SAMARA TRILLING

Mortgage lenders increasingly use machine learning (ML) algorithms to make loan approval and pricing decisions. This has some positive effects: ML loan models can be up to 40 percent less discriminatory than face-to-face lending. But these decisions also present challenges: when ML models discriminate, they do so disproportionately against borrowers without access to mainstream banking services. These borrowers are more likely to be women and people of color. In addition, it is often unclear how existing fair lending laws should be applied to algorithms, and the lending models are updated too frequently for traditional fair lending audits to monitor.

To address these challenges, this project recommends that state lending regulators develop a rigorous definition of statistical fairness for automated lending models. To apply this metric, they should pilot a lightweight, immediate fairness test capable of evaluating ML lending models for compliance. And to test both of these innovations, they should sponsor a Fair Machine Learning Lending Model Contest that is open to the public.

PROBLEM:
EXISTING FAIR LENDING AUDITS ARE INSUFFICIENT

Fair lending audits are valuable tools for detecting and mitigating algorithmic discrimination. But machine learning models that make loan approval and pricing decisions can be updated on a daily basis—far more frequently than traditional fair lending audits are designed to handle. In addition, because ML models lack a human decision-maker, it can be unclear how key concepts of fairness, like discriminatory intent, should be applied to ML algorithms. As a result, traditional fair lending audits are at risk of losing their ability to effectively detect and diagnose discriminatory lending practices.

Fairness Definitions Are Hard To Codify

Mortgage lending and pricing decisions are increasingly made by algorithms: about 5.5 percent of mortgages are granted by algorithms, and this number is growing. Furthermore, a 2019 study from researchers at the University of California, Berkeley found that financial technology (fintech) mortgage lending algorithms discriminated 40 percent less than face-to-face lenders, but failed to eliminate all illegal discrimination. This means algorithms have potential to improve fairness outcomes for borrowers, but currently continue to discriminate against protected classes. This is in part because current legal fairness definitions are subjective enough that they cannot be reliably coded into automated models.

It Is Unclear How to Apply Legal Precedent

There are many ways to calculate fairness for algorithmic loan decisions. For housing loans, the two relevant legal fairness concepts are disparate treatment and disparate impact. Disparate treatment is defined as negative treatment of a loan candidate or group of loan candidates due solely to their protected status (race, ethnicity, gender, etc.). Disparate impact is defined as the unintentional but systemic negative treatment of a protected group of loan candidates. Because ML models lack a human decision-maker to ask about intent or reasoning for a lending decision, it is not always clear how regulators and banks should interpret these legal concepts in statistical settings.

Lack of Clarity Creates Disincentives for Action

The lack of clarity in fairness definitions results in banks spending significant amounts of money on internal risk and fair lending software without guaranteeing a fairer result. Furthermore, this uncertainty creates a disincentive for banks to investigate and improve the fairness of their existing models. Because banks are unsure of how fair lending guidelines will be applied to algorithms, it is easier for banks to avoid liability by not investigating the fairness of their own models and pleading ignorance should issues arise.

SOLUTION 1
DEVELOP A STATEWIDE FAIRNESS DEFINITION

Given these obstacles, state lending regulators should aspire to develop a clearer definition of fairness that is easily applied to machine learning algorithms. Such a definition should be quick and automatable; a human should not need to be present to conduct the audit.

If such a definition has already been developed and vetted by state banking regulators, it should be immediately released to incentivize action.

If a definition has been developed, but would benefit from real-world testing before it is implemented as a regulatory requirement, state lending regulators should pilot the definition. One mechanism by which to achieve such a pilot is through a fair lending contest. Entrants would develop machine learning lending models that optimize for the new fairness definition. Regulators would analyze the decisions made by all submitted models and decide if the definition properly incentivizes non-discriminatory behavior. Such a contest is described below in Solution 3.

If such a definition does not yet exist, state lending regulators should develop a short list of contender definitions from classic statistical fairness definitions and then use the aforementioned fair lending contest to select one.

Examples of potential fairness definitions include:

Statistical parity: does each demographic group get an equal number of loans?

Conditional statistical parity: does each demographic group get an equal number of loans, conditional on the creditworthiness of the individual?

Equal opportunity: does each demographic group have a similar false negative rate (the number of people in the group who were incorrectly denied a loan)?

Equalized odds: does each demographic group have a similar false negative rate *and* false positive rate (the number of people in the group who were incorrectly given a loan)?

Individual fairness: do two people with similar creditworthiness characteristics get similar loan approval outcomes?

For more information about each of these definitions, with the benefits and drawbacks of each, see this matrix.

SOLUTION 2
LIGHTWEIGHT IMMEDIATE FAIRNESS TEST (LIFT)

Second, to effectively audit fast-changing ML lending models, state lending regulators should pilot a lightweight, immediate fairness test (LIFT) to quickly and easily audit algorithmic mortgage models every time they are updated.

A LIFT[1] is a software program that directly queries a lender's ML model with hypothetical applicant information. It receives loan decisions and prices back from the model, and evaluates the results in aggregate for compliance with fair lending policy. Such a system would allow state lending regulators to audit every version of a lender's algorithm that affects consumers, not just the one in effect at audit time. Using a LIFT would also save money by reducing in-person audits. Finally, because it would allow state lending regulators to query a lending model for its future behavior, a LIFT would enable the agency to more effectively prevent, rather than simply punish, discrimination.

There is robust precedent in software engineering for this type of automated testing. Software products can change daily as bugs are fixed, systems are upgraded, and new features are added. With a practice called continuous integration, every time a piece of code changes, a set of tests are automatically run on the entire system to make sure the code does not break existing functionality or have any unintended consequences that degrade the system's security, robustness, or user experience. Until these tests pass, the code is not deployed to users. The tests usually take between a few seconds and 15 minutes to run.

The same concept could be used for fair lending. For each lending model update, a set of fair lending tests must run and pass before the updated lending model is deployed to evaluate real borrowers.

1 *LIFT is a new coinage, not an existing industry term.*

Implementation

Once a definition of statistical fairness is agreed upon, a LIFT could be written to measure approvals, prices, and error rates for demographic groups. There are two options:

(1) State lending regulators could provide this test for banks to run voluntarily, with the results reported back to them; or

(2) State lending regulators could require banks to provide continuous access to the application programming interface (API) for their models so that regulators could query the model and run tests at a chosen cadence.

State lending regulators could then provide counsel or take regulatory action against firms that repeatedly fail fairness tests.

Remedying Discrimination

It is important that this fair lending test also articulate what lenders should do next if they are not meeting requirements. In the past, statistical analysis tools lacked the power to propose less discriminatory alternative models. However, recent interviews with fair lending compliance consulting firms like ZestAI and BLDS suggest that alternative model suggestions are now a staple of their service offerings. State lending regulators could provide instructions on how to bring a model into compliance (highlighting variables with high correlation to protected classes, recommending additional data sources that might provide better accuracy among uncommon borrowers, etc.) and could recommend that lenders seek additional counsel from other fair lending professionals who use ML models.

SOLUTION 3
FAIR MACHINE LEARNING MODEL CONTEST

Because of the importance and sensitivity of both of the above policy changes, state lending regulators should pilot the fairness definition and LIFT by co-sponsoring a fair ML lending model contest with a local university data science program. The contest would invite participants to build a machine learning mortgage lending model optimized for a given definition of statistical fairness. Models should accept a borrower profile (including the loan amount and several measurements of creditworthiness) and output a loan decision (yes or no) and an interest rate. It should be based on a training data set provided by the contest organizers that contains historic borrower profiles, loan decisions, and loan performance; a fairness definition; and a suite of automated tests that assess the model according to the fairness definition.

While it is possible for state lending regulators to reason about the effects of a particular fairness definition by studying academic literature, the effects of any definition on borrowers in their state are highly dependent upon the size, representativeness, and other qualities of the training and test datasets used. A bad definition could even end up increasing algorithmic discrimination against communities that state lending regulators aim to help. A contest that allows teams to build models on real-world datasets would provide state lending regulators with concrete evidence of the impact of the fairness definition on borrowers. The regulators could then alter the definition if it does not meet their goals.

A contest would also help pilot technical infrastructure for a LIFT without adding significant costs. As surveys suggest, regulatory technology can often be expensive to build. By partnering with a university to develop a prototype of software for the contest, state lending regulators can validate the usefulness of the technology without a large budget.

For more about how to run such a contest, see Appendix: Fair Lending Contest.

CONCLUSION

State lending regulators have a unique opportunity to help Americans by setting proactive goals and guardrails for the soon-to-be-ubiquitous automation of consumer lending. Clearly defining a statistical fairness metric and developing a lightweight immediate fairness test would increase regulatory certainty and incentivize the development of fair lending models. Piloting the fairness metric and fairness test through a Fair Machine Learning Lending Model Contest would provide the tangible technical foundation for a leading regulatory program that would enable better ML lending practices and protect borrowers as they make the major financial decisions.

ASPEN TECH POLICY HUB

POLICY

ABOUT THE HUB

The Aspen Tech Policy Hub is a Bay Area policy incubator, training a new generation of tech policy entrepreneurs. We take tech experts, teach them the policy process, and support them in creating outside-the-box solutions to society's problems.

The Aspen Institute
2300 N St. NW, Suite 700
Washington, DC 20037
202 736 5800

THE ASPEN INSTITUTE

PRECEDENT

In January 2019, FICO partnered with Google and several universities, including UC Berkeley, MIT, and Oxford, to sponsor an Explainable Machine Learning Challenge. Teams used a FICO-provided credit training dataset to create ML models that granted or denied credit applications and adequately explained their decisions. With a $5,000 grand prize, the challenge garnered creative, high-quality submissions. FICO said the results proved that machine learning models could be explainable and interpretable, and the dataset provided by FICO has been used productively in ML fairness research well beyond the contest scope.

This type of competition is expected to become more prevalent as a way to explore the potential of machine learning for a particular application before investing large sums in full system development.

GOALS

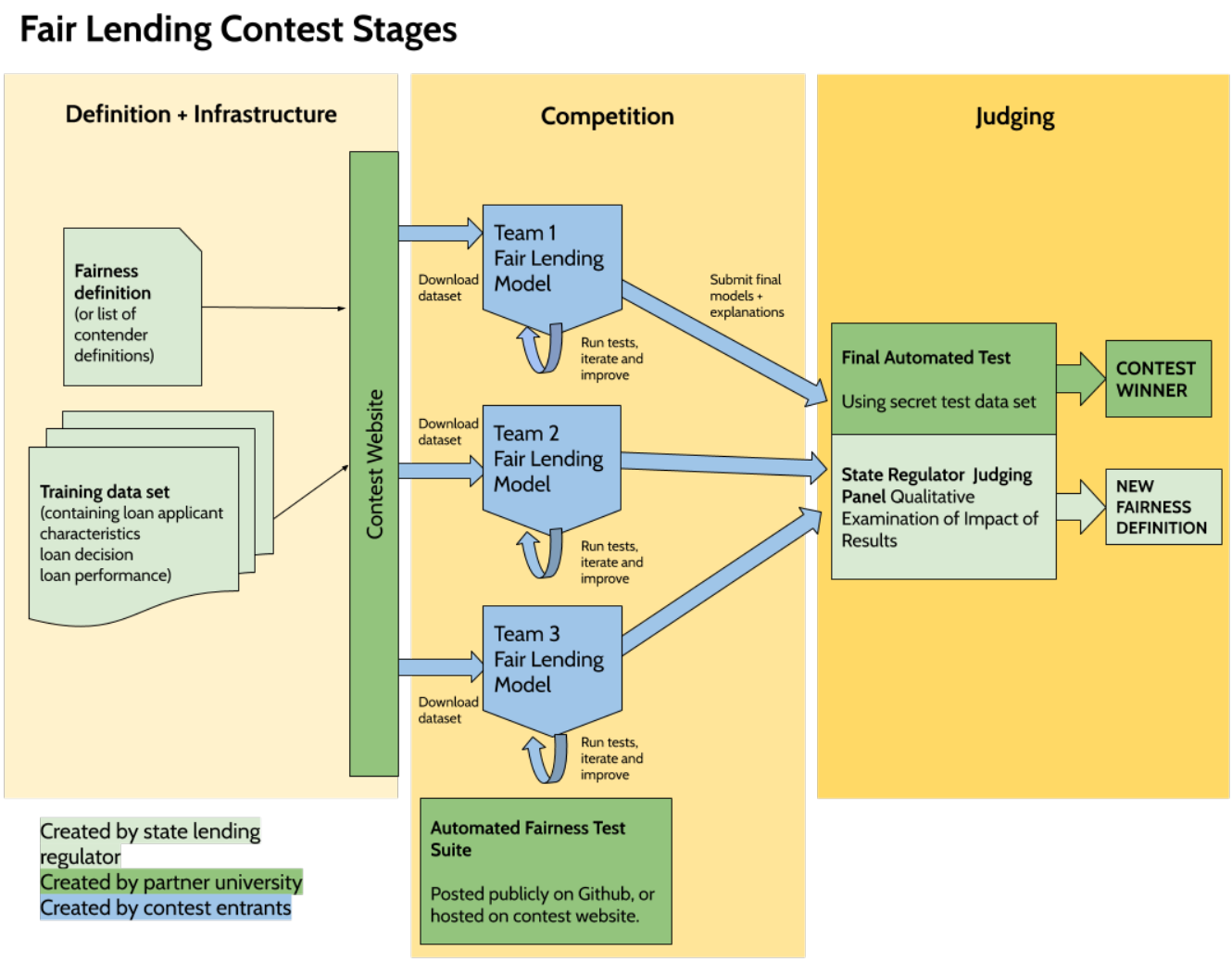
Rather than soliciting examples of explainable machine learning models like the FICO challenge, a Fair Machine Learning Lending Contest would solicit examples of fair models, according to a new automatable definition of fairness. The submitted models would give state regulators insight into which lending decisions might be made under that fairness definition. With this evidence, regulators could confidently decide to adopt the definition or amend it to remedy unintended effects.

In addition, the contest would be judged with a lightweight automated fairness test that scored models on how well they adhered to the new fairness definition. If regulators find this automated fairness test to be useful, they could consider expanding it for regulatory use, to evaluate supervised lenders' machine learning models before giving a stamp of approval.

IMPLEMENTATION

The details for the fair lending machine learning model contest depend on whether state lending regulators already have a chosen definition for ML lending fairness or are deciding between several. If regulators have a strong hypothesis about which fairness definition they'd like to test (e.g., statistical parity or equalized odds), all contestants could be asked to optimize for that definition. If the regulators are choosing between two definitions, they could allow applicants to choose between the two. No matter which definition teams optimize for, it would be easy to test each model with regard to all of potential fairness definitions.

Below is a diagram illustrating a sample fair lending contest's structure and stages:



Fairness Definition

If state lending regulators have a definition ready, they would provide all entrants with:

a. The definition of fairness, in plain language and with several examples of fair and unfair algorithms
b. An automated fairness test that evaluates the model on the fairness definition.

If choosing between several definitions, state lending regulators would provide all entrants with:

a. A list of the fairness definitions under consideration.
b. A suite of automated fairness tests that evaluate a model on all the fairness definitions regulators are considering adopting.

In both cases, regulators would provide applicants with a sample training data set that includes historical state Home Mortgage Disclosure Act (HMDA) data (scrubbed of details about race, gender, and other protected characteristics) joined to loan performance data. HMDA data comprises national loan-level mortgage data collected by banks and made publicly available to aid fair lending investigations. This dataset should include all applicant characteristics that are reported as part of HMDA (except the race, sex, ethnicity, and age categories) as well as whether or not each applicant repaid their loan.

Training Data Set

There are two options for creating a training data set:

Option One

Release a dataset that merges state HMDA data (with protected class information removed) with loan performance data from credit bureaus or government-sponsored enterprises like Fannie Mae and Freddie Mac.

Using a real dataset would provide the clearest prediction of how each fairness definition would affect borrowers.

This dataset should include, at minimum:

- Amount of loan
- Loan-to-value ratio
- Debt-to-income ratio
- Loan approval or denial
- Loan interest rate
- Whether the loan was delinquent for more than a certain number of weeks over the course of its life

Why these fields?

Providing a dataset with financial fitness indicators from HMDA – like loan-to-value ratio and debt-to-income ratio – coupled with the performance of the loan, would allow entrants to train fair lending models without needing access to protected class information like race and gender.

It is important to include loan performance data in addition to loan approval data in this training data set so that models can be trained to predict loan performance directly. Loan approval is not a direct proxy for loan performance, and treating it as such risks perpetuating existing biases in loan approval and pricing processes.

What are the risks?

The risks of releasing this combined dataset are small.

One risk is that, even though the released dataset will not include protected class information, it is possible to attempt to match the HMDA columns of the contest dataset to public HMDA records in order to discover the protected class fields. This does not pose a threat to the integrity of the contest, since having access to this additional protected class information would not meaningfully help contestants. The two most salient concerns are a potential risk to privacy (as linking these two datasets could make it easier to re-identify a borrower or learn that an identified borrower defaulted on a loan); and that, depending on the trends visible in the data, politically motivated actors could use the link between protected class information and loan default data to draw misleading conclusions about which demographic groups default on loans more frequently.

Merging HMDA and loan performance datasets (including protected characteristics) is already possible – researchers have done so using a matching algorithm – so releasing this more limited dataset publicly carries little risk from a privacy perspective.

Regarding generalizations about demographics, data about the correlation between demographic group and repayment rate is already widely available and well explained: historical discrimination can contribute to the financial precarity that leads to loan default.

Nevertheless, to mitigate these risks, state lending regulators should review the trends present in the dataset before publication to address potential incorrect conclusions. Regulators should also consider choosing HMDA fields with the lowest risk of borrower re-identification for inclusion in the dataset. The fewer HMDA fields are included in the dataset, the smaller the risk. Of course, including fewer fields also limits the potential accuracy of a lending model trained on these fields, so these objectives must be balanced.

Option Two

Work with university partners to create a fake dataset that is similar in distribution to the HMDA data and loan performance data, but without any real values.

A fake dataset would carry zero risk of borrower re-identification because the data is fabricated. However, it requires far more work to develop a fake dataset that represents a real community. Moreover, there is a chance that performing well on this fake dataset will not translate to performing well on real datasets, and so this dataset would provide less certainty about how a fairness metric would affect the state borrower community.

FAIRNESS TEST

If state lending regulators have a fairness definition ready, they should work with the partner university to write a fairness test (a LIFT) for that definition of fairness. The fairness test would query a lending model with a predefined set of potential applicants with known demographic information and loan outcomes, and conduct statistical analysis on which loans were granted and at what price. The test would vary based on the fairness definition, but many fairness tests can be simple measurements of approval rates and error rates for different demographic groups.

State lending regulators would require entrant teams to submit a pull request containing their ma-

chine learning model to the contest's GitHub repository. This fairness test would be available as a continuous integration test. This means that entrants could run the fairness test over their model as many times as they like during development, iterating until the model performs well on the test.

If state lending regulators are choosing between several definitions, lending regulators would write fairness tests for each of the candidate definitions, and require entrant teams to select one definition of fairness for which to optimize. Regulators could then award a prize to the team that performs best on each fairness definition.

JUDGING AND PRIZES

The competition period should last for 3–5 months. An automated fairness test, similar to the one provided to contest entrants, should be used to judge the contest, but state lending regulators should use a different, secret set of test data for the judging test. This would prevent models from "over-fitting," performing well on the public LIFT test data but not on any other legitimate data sets. The model that performs best on the fairness definition should be declared the winner.

Similar competitions sponsored by Kaggle and FICO have given top prizes of $5,000. A similar sum, administered through the university partner, could be awarded here.

Separately, state lending regulators should use the performance of the submitted models to analyze the effects each fairness definition would have on lenders and borrowers in their state. If regulators have a fairness definition ready, this analysis would help uncover any unintended consequences of the proposed definition (e.g. certain demographic groups negatively impacted) and allow regulators to amend the fairness definition before making it official. If state lending regulators are choosing between definitions, this analysis would allow regulators to select the fairness definition they feel would best serve their constituent communities.

For more detail, see the full set of proposed contest rules.

ASPEN TECH
POLICY HUB

Fair Algorithmic Housing Loans

THE ASPEN INSTITUTE