ASPEN TECH POLICY HUB

THE ASPEN INSTITUTE

Scott Hallworth
Chief Data, Model & Analytics Officer
Fannie Mae
1100 15th St. NW
Washington, D.C. 20005

Frank Nazzaro
Chief Information Officer
Freddie Mac
8200 Jones Branch Dr.
McLean, VA 22102-3110

July 8, 2020

Dear Mr. Hallworth and Mr. Nazzaro,

As a computer scientist, researcher, and fellow at the Aspen Institute's Tech Policy Hub, I commend and thank you for Fannie Mae and Freddie Mac's commitments to data transparency. The single family loan-level performance datasets you have released to the public are wonderful examples of effective open data for the public good.[1]

I write to request, on behalf of the group of researchers who have signed the attached letter, that **Fannie Mae and Freddie Mac publicly release a combined dataset merging borrower characteristics and loan-pricing data with the existing single-family mortgage loan performance dataset.** Such a dataset will lower the cost of building fair machine-learning (ML) lending models and increase the models' accuracy.

This brief letter provides background on algorithmic mortgage lending, outlines why existing data is insufficient, and explains the need for a new combined dataset that will help correct the deficiencies.

BACKGROUND

Mortgage lending and pricing decisions are increasingly made by algorithms: about 5.5 percent of mortgages are decided by algorithms, and this number is growing.[2]

Historical loan performance data is crucial to the development of these algorithms. ML lending models learn to predict future loan performance by examining paired examples of borrower credit characteristics and the resulting loan performance. New financial technology ("fintech") lenders often use datasets like the loan performance dataset you have released to train their new ML lending models.[3]

However, because credit characteristics often correlate with protected attributes such as race and gender, ML lending models can discriminate against protected groups, even if the protected attributes, such as race or gender, are not included in training or test data. In order to determine whether an algorithm discriminates against a certain group, it is necessary to test the algorithm by feeding it with borrower profiles with known race and gender, and comparing how often it grants loans to legally protected groups versus unprotected groups.

LOAN APPROVAL DATA (e.g. HMDA) IS INSUFFICIENT TO TRAIN ALGORITHMS

The best way to train an ML model to predict loan default is to provide examples of applicants' characteristics and whether they ended up repaying their loan. Unfortunately, most of the loan performance datasets used to train and test algorithms do not include race and gender data, and so cannot be used for fair lending analysis. Similarly, the Home Mortgage Disclosure Act (HMDA) datasets that do include race and gender data do not include loan-level performance. Thus, there is currently no single dataset that adequately supports both building ML lending models and evaluating them for fairness.

Because datasets containing matched pairs of applicant characteristics and loan performance are not readily available, a common shortcut is to use pairs of applicant characteristics and loan approval and price (i.e., public HMDA datasets), instead of loan performance. This assumes that loan approval and price are direct proxies for loan performance, which they are not. More worryingly, this approach risks perpetuating existing biases in loan approval and pricing processes.

For example, an applicant may be judged to be high-risk by a ML risk-based pricing system, and thus would be offered a high interest rate. If that applicant accepts the mortgage and repays it without issue, that provides valuable information about the accuracy of the algorithm that initially tagged them as high risk. We would call this a false positive: the system predicted the applicant would be at high risk of default, but in reality they were low risk.

Similarly, an applicant who is judged to be low-risk — and offered a lower interest rate — ultimately may not repay that loan. This would suggest a false negative: the system determined the applicant to be low-risk when they should have been tagged as higher risk.

Without a dataset that matches applicant characteristics to loan performance, an ML model would not be able to learn when the loan-pricing system was wrong in assessing the applicant's risk. Loan-performance data is crucial for improving the accuracy of risk-based pricing ML systems over time, and for giving consumers interest rates that accurately describe their risk of default.

There are workarounds for merging applicant characteristics and loan performance – such as purchasing credit bureau datasets that include variables closely correlated with race and gender, or using algorithms to match datasets that include race and gender (e.g., HMDA data) to loan performance datasets[4] – but these methods are expensive and prone to error.[5] A mismatch could mean that an algorithm learns an incorrect relationship between borrower characteristics and loan repayment, making it more likely that the algorithm will make incorrect loan decisions.

If ML model developers had access to a reliable dataset that included:

a. the borrower characteristics reported under HMDA
b. the loan decision and price, and
c. the loan performance data points from the existing GSE mortgage performance datasets

they would be able to build models that more accurately predict default, and analyze and correct their models for unfair bias.

PRIVACY CAN BE PRESERVED WITH A MERGED DATASET

One potential concern with the release of a dataset that merges borrower characteristics with loan performance is the risk of borrower re-identification. While

this is an important and valid consideration, it is unlikely that inclusion of loan performance in a merged dataset would significantly increase the risk of re-identification, for two reasons.

First, as mentioned above, it is already possible to algorithmically match loan performance and borrower characteristics across datasets, just with a lower level of accuracy than an official dataset would provide.[6] The most significant concern is that an official dataset would increase the accuracy of matches, which would give a higher level of confidence to borrower re-identification. However, we believe that this potential risk is balanced by the benefits that more accurate ML models would provide – namely, fewer bad loan decisions that could systematically harm borrowers.

Second, loan performance data by itself would not help identify a borrower across datasets. For context, similar privacy questions were explored in the Bureau of Consumer Financial Protection's 2018 rule regarding HMDA data disclosure.[7] The Bureau declined to include the unique loan number known as a universal loan identifier (ULI) in public HMDA data because of concerns about borrower re-identification. This position is understandable: because a ULI uniquely identifies an individual loan, it could be used to identify that same loan in other data sets. Loan performance, however, does not uniquely identify a loan (especially if it is published as a true/false value indicating whether the loan was delinquent for over 90 days), and hence is less sensitive.[8]

In order to lower the cost and increase the accuracy of fair machine-learning lending models, I respectfully request that you release a combined dataset that merges HMDA borrower characteristics and loan-pricing data with the existing single-family mortgage loan performance data.

Thank you very much for your consideration.

Samara Trilling
Fellow, Aspen Institute Tech Policy Hub

Notes:

1. http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page
   https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html
2. "The trend for most mortgage lenders is clearly toward algorithmic decision-making." https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf
3. Based on conversations with leaders in fintech AI lending who reported that high-quality, publicly available data like Fannie Mae's and Freddie Mac's allowed them to bootstrap their lending models to start their businesses.
4. "Since there are no unique mortgage loan identifiers in the U.S., we develop an algorithm using classifier techniques to match loans found in two independent datasets: the McDash dataset, which contains loan-level data compiled by Black Knight Financial Services, and the ATTOM dataset...Our merging process applies a modified k-nearest-neighbor classifier". https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf
5. Researchers interviewed quoted credit bureau data as costing thousands to tens of thousands of dollars. While this could be seen as table stakes for a new fintech lender, researchers often have much smaller budgets. The presence of a financial hurdle to even exploring the dataset is enough to deter the creation of a new model, as I experienced as a fellow attempting to create my own model at the Aspen Institute's Tech Policy Hub.
6. https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf
7. https://files.consumerfinance.gov/f/documents/HMDA_Disclosure_FPG_--_Final_12.21.2018_for_website_with_date.pdf
8. FICO used 90-day delinquency as the boolean value for a "bad" loan in its explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2