



WHITE PAPER· By Graydon Goss

Balancing Fairness, Accuracy, and Explainability in Machine Learning Models for Credit Decisioning

As machine learning (ML) reshapes the landscape of credit decisions, the stakes around lending fairness become much more than an ethical concern. When fairness is overlooked in credit decisions, financial institutions risk facing regulatory penalties, and losing trust from stakeholders and borrowers. Mitigating these risks requires that models balance accuracy, fairness, and transparency – a challenging task given that enhancing one trait often comes at the cost of the others. How, then, can financial institutions create models that are explainable, accurate, and fair all at once?

To answer this question, let's first gain a better understanding of the current state of lending fairness. FI Consulting (FI) analyzed recent data collected under the Home Mortgage Disclosure Act (HMDA), which requires many financial institutions in the U.S. to disclose information on home mortgage applications.

This dataset has over 14 million home mortgage applications from 4,460 institutions, providing sufficient data for the analysis of trends across individual financial institutions and the entire U.S. home mortgage market.

These trends were captured on an interactive dashboard. The gauge charts at the bottom of Figure 1 show fairness scores according to the HMDA data, including scores for age group, race and gender for each financial institution. To calculate these benchmark scores, FI created a robust scoring method by combining three concepts of ML fairness: Adverse impact ratios (AIRs), demographic parity, and equal opportunity.



Figure 1. Fair Lending Analysis for Individual Financial Institutions

- **Adverse Impact Ratios:** AIRs are a standardized measurement of fairness. They are calculated by dividing the preferred treatment rates for each demographic subgroup by the preferred treatment rate of the best-off group. For example, if men receive loans 80% of the time and women receive loans 60% of the time, the AIR for men would be $80\%/80\% = 100\%$, and the AIR for women would be $60\%/80\% = 75\%$.
- **Demographic Parity:** Demographic parity is a concept which states that preferred treatment rates should be the same across all demographic subgroups. For example, if men receive loans 80% of the time when they apply, then women should receive loans at a rate of 80% too.
- **Equal Opportunity:** Equal opportunity is a concept which states that the true positive rates (TPRs) should be the same across demographic subgroups. The TPR is the ratio of individuals that rightfully received preferred treatment to all observations that deserve preferred treatment. For example, equal opportunity applies if both men and women receive loans 90% of the time when they rightfully should.

$$\text{true positive rate} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

The dashboard visualizations show that there was a non-negligible trend of lending discrimination (lack of fairness) that existed in the U.S. home mortgage market. Now that there is proof the model-building process needs to be improved, it is necessary to revisit the question “How can organizations create an explainable model that optimizes accuracy without being unfair?” To address this, FI built an explainable, predictive Machine Learning (ML) model with the HMDA data, then tested if it was possible to increase fairness without sacrificing accuracy. FI found that improved feature selection and hyperparameter tuning best increased model fairness (AIR) with negligible impact on model performance (AUC). The results for a single demographic are shown in the following figure.

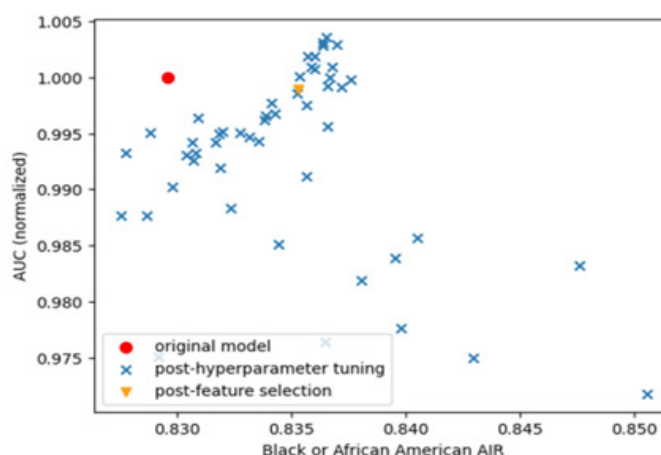


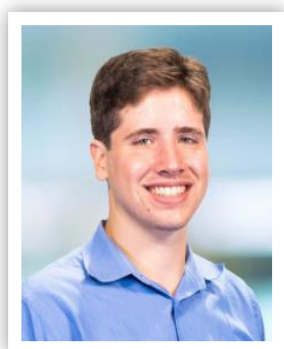
Figure 3. Accuracy and Fairness Scores after Post-Processing Techniques

While feature selection and hyperparameter tuning worked in this case, it is not a “one size fits all” solution. When tuning your credit decisioning models, it is important to have a strong sense of the data you are working with to avoid overfitting. It is also important to select intuitive features with strong business sense, else risk creating an accurate model that is not explainable.

Takeaways

Leveraging ML solutions is critical for financial institutions to remain competitive in today’s landscape, but inherently discriminatory and unexplainable ML models risk both financial penalties and harm to a company’s reputation. Keeping models fair, explainable, and accurate is a difficult needle to thread, but is essential for ensuring trust in the models’ use and operation.





Author: Graydon Goss

Gray is a consultant at FI Consulting, specializing in analytics, modeling, and machine learning. He has supported financial institutions and federal clients by developing entity resolution frameworks, graph-powered analytics, GenAI chatbots, and end-to-end ETL pipelines. He earned dual bachelor's degrees in Applied Statistics and Economics from the University of Virginia.

To learn more about how FI Consulting created fair and accurate ML models and communicated the state of the home mortgage market through visual, quantitative, and statistical analysis, please email modeling@ficonsulting.com.

FI Consulting can help improve financial institutions' predictive modeling systems and create dashboarding tools to extract novel insights from complex datasets.



FI•CONSULTING

2107 Wilson Boulevard | Suite 400 | Arlington, VA 22201 | 571.255.6900 | ficonsulting.com | info@ficonsulting.com

© 2026 FI Consulting. All Rights Reserved.