Selecting For Less Discriminatory Algorithms:

A Relational Search Framework for Navigating Fairness-Accuracy Tradeoffs in Practice

Hana Samad,[1] Michael Akinwumi,[1,a] Jameel Khan,[2] Christoph Mugge-Durum,[1] and Emmanuel O. Ogundimu[3]
[1]Responsible AI Lab, National Fair Housing Alliance,[2] Housing & Community Development, National Fair Housing Alliance, Washington DC,[3]Department of Mathematical Sciences, Durham University, Durham, United Kingdom
[a]makinwumi@nationalfairhousing.org

Abstract

As machine learning models are increasingly used for high-stakes decision-making, challenges arise in selecting a fair model, particularly in lending decisions where it is a civil rights obligation. Rather than optimizing fairness within a single model, this work tests **horizontal LDA search**—comparing fairness across model families to identify *Less Discriminatory Algorithms (LDAs)*. Drawing on the concept of **model multiplicity**, which recognizes that models with similar accuracy can yield different fairness outcomes, we treat model selection itself as a crucial step in making fairness gains. Using 2021 HMDA data, we extend Lee and Floridi's relational fairness framework to demonstrate that early-stage model selection across families can achieve appropriate fairness gains. This lightweight approach of using a relational fairness framework to evaluate models across families presents a context-aware solution to balancing fairness and accuracy, even under real-world resource constraints.

Methods

Dataset:
2021 Home Mortgage Disclosure Act (HMDA) data to simulate U.S. mortgage lending outcomes, modeling loan approval and denial to evaluate fairness and financial inclusion.
Machine Learning Models:
1) Logistic Regression (LR) 2) K-Nearest Neighbors (KNN) 3) Classification and Regression Tree (CART) 4) Gaussian Naïve Bayes (NB) 5) Random Forest (RF)
Fairness Metrics:
Equal Opportunity (EOP), Positive Predictive Parity (PPP), False Positive Error Rate Balance (FPERB), Demographic Parity (DP), Equalized Odds (EO)
Analytic Framework:
Applied *Lee & Floridi (2021)*'s **relational trade-off framework** to test Horizontal LDA Search method and visualize fairness–accuracy trade-offs across models, comparing financial inclusion (expected loan value) vs. negative impact (denial rate for Black applicants).
Horizontal LDA Search:
Compares *across* model families to identify *Less Discriminatory Algorithms (LDAs)* that balance fairness, accuracy, and inclusion at the model-selection stage. Compare to **Vertical LDA Search:** Tunes hyperparameters *within* a single model family to optimize fairness.

Acknowledgements: The Responsible AI Lab (RAIL) at the National Fair Housing Alliance (NFHA) would like to sincerely thank Wells Fargo for supporting this ongoing research.

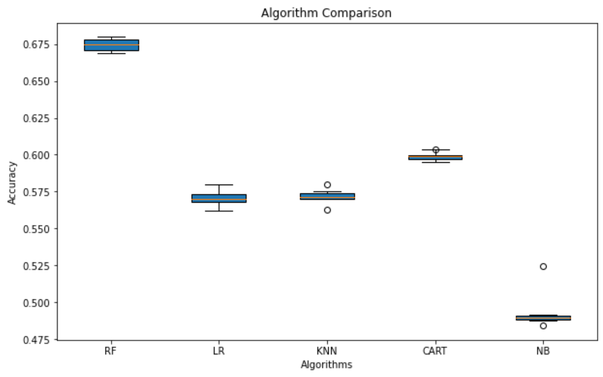
Approaches to Fairness Testing

Table 1: Assessing Equity: A Comparative Analysis of Fairness Metrics Between Non Hispanic White and Black Applicants.

Fairness Metric / Model	Equal Opportunity (EOP)	False Positive Error Rate Balance (FPERB)	Equal Odds (EO)	Positive Predictive Parity (PPP)	Positive Class Balance (PCB)	Negative Class Balance (NCB)
LR	6%	10%	6%	18%	2%	2%
KNN	5%	4%	5%	19%	4%	2%
CART	11%	17%	11%	14%	11%	17%
NB	14%	5%	14%	18%	9%	6%
RF	24%	43%	24%	3%	13%	18%

Legend: The table compares the fairness performance of five ML models – LR, KNN, CART, NB, and RF across six fairness metrics calculated as percentage disparities between white and black applicants. Lower percentages indicate smaller group disparities. The disparity in performance per race indicates inherent structural differences between model families. Horizontal LDA search as proposed in this paper, along with the relational tradeoff framework, exploits these structural differences to surface variable fairness effects

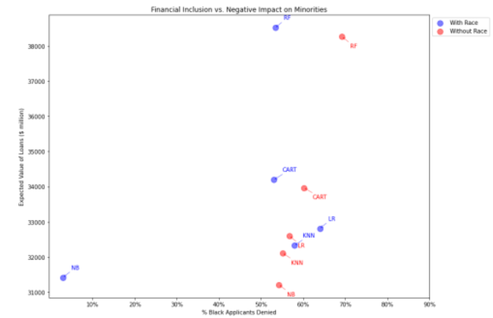
Figure 1: Algorithm Accuracy without Demographic Features of Race, Sex, and Minority Population of the Census Tract Area



Legend: The Box Plot representing the accuracy of each ML algorithm. Accuracy is the proportion of all classifications that are correct calculated as (TP + TN)/(TP + TN + FP + FN). While no single algorithm guarantees optimal fairness, leveraging the concept of model multiplicity offers a promising path forward, demonstrating that multiple model families can achieve comparable predictive accuracy.

The Relational Tradeoff Framework

Figure 2: Trade-off: financial inclusion vs. negative impact on minorities— with and without Race



Legend: This scatterplot compares the relationship between expected loan value and denial rates of black applicants across out five ML model types. The figure highlights that including race as a feature improves both fairness (lower denial rates) and inclusion (higher loan value) underscoring the potential for race-aware modeling to mitigate adverse impact.

Results and Horizontal LDA Search

- Including race as a feature improved financial inclusion (expected loan value) and reduced denial rates for Black borrowers, showing that fairness interventions can enhance outcomes for minority borrowers.
- The horizontal LDA search framework enables practitioners to compare fairness–accuracy trade-offs *across* model families, helping identify less discriminatory algorithms without heavy computational cost.
- Horizontal LDA search provides a resource efficient method for lenders to make fairness gains early in their selection process for a ML model.
- Fairness and accuracy need not be at odds; selecting the *right* model family early in development can yield both equitable and effective outcomes.

Full Paper

Website

Durham University

NFHA NATIONAL FAIR HOUSING ALLIANCE