Fair A.I. for Mortgage Approvals: Balancing Accuracy and Equity

David Hill
Department of Mathematics & Statistics
York University
1davidhill@gmail.com

Elaheh Zarabi
Department of Economics
York University
elizrb72@yorku.ca

Rishigesh Patgunarajah
Department of Mathematics & Statistics
York University
rish2250@my.yorku.ca

Vincent Sham
Department of Electrical Engineering & Computer Science
York University
shamvinc@yorku.ca

Zohaib Ahmed
Department of Information Systems & Technology
York University
zohaibabdullah1999@gmail.com

Motivation

The increasing use of machine learning in home mortgage decision-making raises critical concerns around fairness and bias, particularly regarding racial bias in home mortgage approvals.

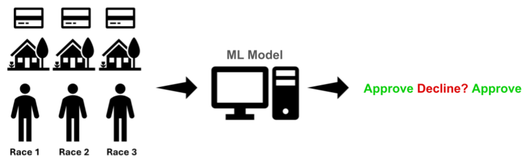

Figure 1. How an AI model makes mortgage decisions across different demographic groups

- Algorithmic decision-making can **unintentionally replicate historical biases**, especially those related to race.
- There is evidence that **80% of Black applicants are more likely to be denied mortgages than white people** by using A.I. systems.[1]
- Financial institutions must ensure that A.I. systems are not only accurate but also fair.

Contributions

By making fair mortgage decisions without sacrificing performance, our key contributions are:

- **Rigorous Feature Selection** to identify the most relevant variables
- **Fairness-aware Hyperparameter Optimization** to balance predictive accuracy and equal opportunity
- **A Generic Resampling Method** that reduces unfairness during training while preserving data integrity to maintain the model performance

Our method can **maintain the accuracy** while **minimizing the bias**. It even achieves higher business value than unfair models. Our method can encourage financial institutions to consider **the adoption of ethical A.I.** in home mortgage lending.

Data

FFIEC Home Mortgage Disclosure Act

- Source: Federal Financial Institutions Examination Council (FFIEC), Home Mortgage Disclosure Act (HMDA).
- Data selected: 2018-2023, State of California
- 13 million records, 99 features
- Key Features: Loan amount, Loan to Value Ratio, Income, Debt to Income Ratio, Loan Term
- Target: Action Taken

Rigorous Feature Selection

To ensure our model makes decisions based on applicant-relevant factors, we implemented a multi-stage feature selection process:

1. Stratified Subsampling
   - Drew a stratified sample of 400,000 applications from over 13 million mortgage records in the HMDA dataset.
   - Used stratified subsampling to Preserve race population ratio and approval ratio for each race.
   - Improved computational efficiency significantly while maintaining a demographically representative training set.
2. Initial Filtering
   - Removed features with over 90%+ missing values.
   - Excluded high-risk data leakage variables (e.g., interest rate, denial reasons, origination charges, etc.) that will not be available in production.
3. Backward Greedy Selection
   - Selected the most important features in a dataset by starting with all the available features and iteratively removing the least significant ones based on the validation accuracy. The resulting dataset has been reduced to 15 features.

Metrics

To assess our model, we used a combination of performance, fairness, and business metrics:

- Performance Metric
  - **Accuracy** measures overall correctness of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Fairness Metric
  - **Equal Opportunity** ensures that good candidates should have the same chance to get approved independent of their race (i.e. $TPR(unprotected\ race) = TPR(protected\ race)$).

$$Fairness\ Score = \frac{1}{1 + |TPR(unprotected\ race) - TPR(protected\ race)|}$$

- Business Metric
  - **Estimated Business Value** measures the values of deploying ML models for mortgage approval.

$$B = mortgage\ amount \times interest\ rate \times mortgage\ term + origination\ charges - default\ rate \times mortgage\ amount$$

  - Assumptions - (1) TP - Good default rate = 1.0% (2) FP - Bad default rate = 4.0% (3) Constant interest rate (4) Race-sensitive median for missing values

Our method

Our baseline model is Light Gradient Boosting Model (LGBM), which contains multiple hyperparameters. In conventional machine learning workflows, hyperparameters are tuned solely to maximize predictive performance. However, the resulting model cannot ensure fairness.

Fairness-aware Hyperparameter Optimization

Fairness can be achieved through the hyperparameter optimization with the **dual objective of accuracy and fairness score** [2]. Tree-structured Parzen estimator (TPE), a Bayesian optimization method, and Hyperband Pruner (TBP) are used to speed up the process of hyperparameter optimization.

A Generic Resampling Method

In Fairness and Bias literature, resampling methods are adopted to ensure the balance and fairness of all races in the training dataset [3]. However, it hurt the performance by putting too much emphasis on the protected group. Therefore, we propose a generic resampling method to **adjust the resampling ratio dynamically**.

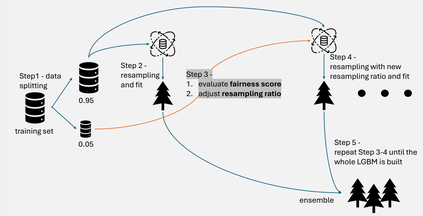

Figure 2. High-level overview of data pipeline.

Results

This plot compares hyperparameter optimization strategies, showing that our method achieves a strong balance between high performance and fairness. Our method plus hyperparameter optimization with performance only form a frontier for **the best performance-fairness tradeoff**.

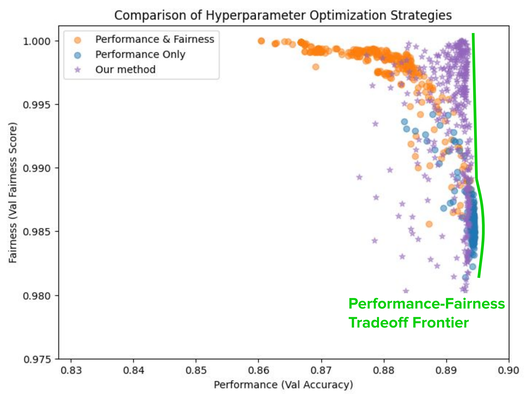

Figure 3. Hyperparameter tuning plot displaying performance–fairness trade-offs.

According to table 1, our method can maintain performance while minimizing racial bias. Most importantly, it even has a higher business metric than performance focused models. Hence, this shows that our method achieves an optimal balance between accuracy and fairness—lying along the performance-frontier and outperforming performance focused models in key business metrics.

Models	Test Acc	Test FairScore	Test TP	Test FP	Business Metric
Fair	0.860	1.00	323600	50088	$408242
Balance	0.891	0.996	319741	37615	$422453
Performance	0.895	0.987	318421	34714	$425675
Our method	0.894	0.999	316835	33661	$426710
Ground Truth (Perfect Prediction)	-	-	324774	0	$469460

Table 1. Model performance and fairness metric comparisons.

Model Analysis

- **Feature Importance** ranks features by their overall impact on the tree model.

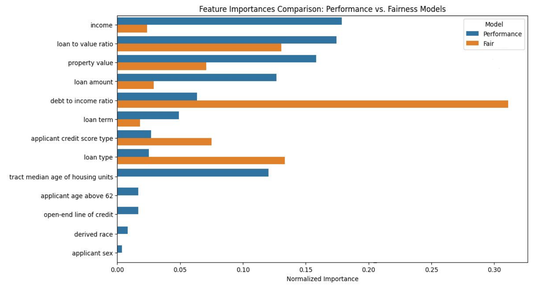

Figure 4. Normalized feature importance comparison of Performance Focused Mod (PFM) vs. Fairness Enhanced Model (FEM).

- **SHAP (SHapley Additive exPlanations)** determines each feature's contribution to an individual prediction.

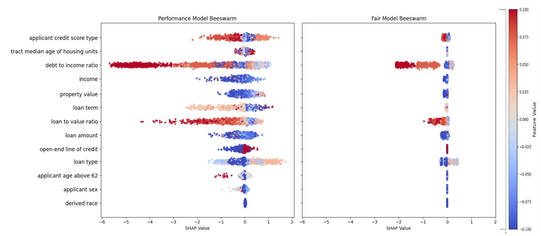

Figure 5. SHAP value feature comparison: PFM vs. FEM.

References

[1] Hale, K. (2021) https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/

[2] Cruz, A. F., Saleiro, P., Belem, C., Soares, C., Bizarro, P. (2021). *Promoting Fairness through Hyperparameter Optimization*. In *IEEE Int. Conf. on Data Mining (ICDM)*, pp. 1036–1041. https://doi.org/10.1109/ICDM51629.2021.00119

[3] Iosifidis, V., Fetahu, B., Ntoutsi, E. (2020). *FAE: A Fairness-Aware Ensemble Framework*. https://arxiv.org/abs/2002.00695

[4] Watanabe, S. (2023). *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*. https://arxiv.org/abs/2304.11127

[5] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A. (2018). *Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization*. arXiv:1603.06560. https://arxiv.org/abs/1603.06560