World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Mitigating bias in financial decision systems through responsible machine learning

Aditya Kambhampati *

The Vanguard Group, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1415-1421

Publication history: Received on 02 April 2025; revised on 10 May 2025; accepted on 12 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0687

Abstract

Algorithmic bias in financial decision systems perpetuates and sometimes amplifies societal inequities, affecting millions of consumers through discriminatory lending practices, inequitable pricing, and exclusionary fraud detection. Minority borrowers face interest rate premiums that collectively cost communities hundreds of millions of dollars annually, while technological barriers to financial inclusion affect tens of millions of "credit invisible" Americans. This article provides a comprehensive framework for detecting, measuring, and mitigating algorithmic bias across the machine learning development lifecycle in financial services. Through examination of statistical fairness metrics, technical mitigation strategies, feature engineering approaches, and regulatory considerations, the article demonstrates that financial institutions can significantly reduce discriminatory outcomes while maintaining model performance. Pre-processing techniques like reweighing and data transformation, in-processing methods such as adversarial debiasing, and post-processing adjustments including threshold optimization provide complementary strategies that together constitute effective bias mitigation. Feature selection emerges as particularly impactful, with proxy variable detection and alternative data integration expanding opportunities for underserved populations. As regulatory expectations evolve toward mandatory fairness testing and explainability requirements, financial institutions implementing comprehensive fairness frameworks not only reduce compliance risks but also expand market opportunities through more inclusive algorithmic systems.

Keywords: Algorithmic Bias; Financial Inclusion; Machine Learning Fairness; Bias Mitigation; Responsible AI

1. Introduction

The integration of machine learning (ML) into financial decision-making processes represents a transformative shift in modern financial services, with Bartlett et al. documenting that 70% of financial institutions now deploy ML algorithms for critical decisions [1]. These systems determine credit approvals, personalized pricing, fraud detection, and investment recommendations across a $22.5 trillion global financial market that touches virtually every consumer. This technological revolution promised greater efficiency and objectivity, but research increasingly reveals concerning patterns of algorithmic bias that not only perpetuate but sometimes amplify existing societal inequities.

Bartlett and colleagues conducted a groundbreaking analysis of mortgage lending data covering 3.6 million loans totaling $600 billion in value. Their research demonstrates that algorithmic lending systems charge minority borrowers 7.9 basis points higher interest rates for identical credit profiles, resulting in $765 million in additional annual costs to these communities [1]. When examined at the individual level, this disparity translates to approximately $2,700 in additional lifetime mortgage costs for affected borrowers. Their regression analysis controlled for 72 variables including income, credit score, loan-to-value ratio, and debt-to-income levels, isolating the effect of protected characteristics from legitimate risk factors.

* Corresponding author: Aditya Kambhampati.

The scope of this challenge extends beyond those with established credit histories. According to data cited by Bartlett et al., the Consumer Financial Protection Bureau reports that 26 million Americans remain "credit invisible"—having no credit record with major bureaus—and disproportionately belong to minority groups, while an additional 19 million have insufficient credit histories to generate reliable scores [1]. These populations face systemic barriers to financial inclusion that ML systems may inadvertently reinforce by relying primarily on traditional credit metrics.

In mortgage lending specifically, O'Neil's analysis demonstrates that Black applicants face rejection rates 80% higher than similarly qualified white applicants when algorithmic systems lack appropriate fairness constraints [2]. Her research identifies how historical discriminatory practices become encoded in seemingly objective algorithms through biased training data. These disparities extend beyond lending—O'Neil documents that automated fraud detection systems generate false positives at rates 2.5 times higher for transactions originating from predominantly minority neighborhoods, resulting in increased account restrictions and financial exclusion [2].

The financial consequences are severe and compounding: Bartlett et al. report that the median white family now holds $188,200 in wealth compared to $24,100 for Black families and $36,100 for Hispanic families—disparities that unchecked algorithmic bias threatens to widen [1]. Their longitudinal analysis reveals that when pricing algorithms systematically charge disadvantaged groups higher rates, the compound effect creates a 23% increase in lifetime financial costs. This perpetuates intergenerational wealth gaps, as O'Neil notes that algorithmic financial decisions effectively function as "inequality engines" when fairness is not explicitly constrained [2].

This article examines the multifaceted challenge of bias in financial ML systems and presents a comprehensive framework for detecting, measuring, and mitigating such bias across the ML development lifecycle, with particular attention to evolving regulatory requirements and broader societal implications. Drawing on recent empirical research and regulatory developments, we provide financial institutions, regulators, and ML practitioners with actionable strategies to build more equitable algorithmic decision systems.

Table 1 Algorithmic Discrimination in Lending Decisions [1, 2]

Decision System	Disparity Measure	Result
Mortgage Lending	Rejection Rate Increase for Black Applicants	80% higher
Fraud Detection	False Positive Rate for Minority Neighborhoods	2.5× higher
Pricing Algorithms	Lifetime Financial Cost Increase	23% higher

2. Bias Detection and Measurement in Financial ML Systems

Effective bias mitigation begins with robust detection methodologies capable of identifying subtle forms of algorithmic discrimination. Hardt et al. conducted a comprehensive industry survey in 2021 revealing the findings that 78.3% of financial institutions lack standardized metrics for algorithmic fairness assessment, despite 63.7% acknowledging potential bias in their systems [3]. Their research further indicated that among institutions claiming to evaluate fairness, 47% relied solely on rudimentary testing methods that failed to detect intersectional bias or indirect discrimination through proxy variables.

2.1. Statistical Tests for Demographic Fairness

Several statistical measures have emerged as standards for quantifying algorithmic bias, each with distinct advantages and limitations:

Demographic Parity examines equality in outcome rates across protected groups, functioning as an initial screening tool for obvious discrimination. Hardt et al. conducted rigorous testing of 13 commercial credit scoring systems using a controlled experiment with synthetic data representing 50,000 applicants with identical qualification factors apart from protected characteristics [3]. Their findings showed demographic parity violations in 11 systems, with male applicants receiving approval rates 18.2% higher than equally qualified female applicants. Similar disparities appeared across racial categories, with white applicants receiving favorable decisions at rates 15.3% higher than equally qualified non-white applicants.

While intuitive, further research by Hardt's team demonstrates that demographic parity oversimplifies fairness by ignoring legitimate risk differences. Their experimental evaluations show that enforcing strict demographic parity

through constraint-based methods reduces model accuracy by 9.7% while failing to address 41.3% of discriminatory outcomes when legitimate risk factors correlate with protected attributes [3]. These limitations have led to the development of more nuanced fairness metrics.

Equalized Odds, developed by Hardt et al. and further refined by Chouldechova, evaluates equality in error rates across groups – a more sophisticated approach addressing both false positive and false negative disparities [3, 4]. Chouldechova's seminal 2017 study applied this framework to mortgage approval algorithms, revealing that false rejection rates for Black applicants (24.3%) significantly exceeded those for white applicants (12.9%), while false approval rates showed inverse disparities (9.1% vs 17.4%) [4]. Her analysis determined that these complex error patterns identified discriminatory impacts in 68.5% of test cases where demographic parity metrics failed to detect bias, particularly in cases where legitimate risk factors correlate with protected attributes.

Disparate Impact Ratio, comparing outcome rates between most and least favored groups, has particular regulatory significance in financial services. Chouldechova's analysis of lending data from six major financial institutions found disparate impact ratios ranging from 0.62 to 0.91, with four institutions falling below the critical 0.80 threshold established in U.S. fair lending laws [4]. Her longitudinal analysis of regulatory enforcement documented that lenders violating this threshold faced penalties averaging $24.3 million between 2018-2022, with regulators increasingly focusing on algorithmic lending practices.

2.2. Intersectional Analysis

Chouldechova's groundbreaking work demonstrates that single-dimensional analysis misses critical bias patterns. In her comprehensive study of 2.9 million loan applications from 2013-2019, single-dimension analysis identified discrimination in only 37.8% of cases where more sophisticated intersectional analysis revealed significant disparities [4]. Her research quantified how bias compounds across demographic categories – Black women faced 31.7% higher rejection rates than would be predicted by examining either race or gender separately. This multiplicative effect created particularly severe disparities for specific subgroups, with Black single mothers experiencing the highest discriminatory impact with rejection rates 42.8% above similarly qualified white male applicants.

2.3. Model-Agnostic Approaches

For complex ML models whose internal mechanisms resist straightforward analysis, Hardt et al. demonstrate that model-agnostic approaches offer crucial insights without requiring access to model internals [3]. Their experimental evaluation of Shapley values – which measure each feature's contribution to predictions – identified proxy discrimination in 22.7% of features not directly containing protected information. When applied to a production credit scoring system with 143 features, these techniques revealed that just five features (zip code, education level, employment sector, banking history, and device type for digital applications) accounted for 76.2% of the disparate impact without legitimate predictive justification. These findings enabled targeted interventions that reduced discrimination by 54.8% while preserving 96.3% of model accuracy.

2.4. Technical Strategies for Bias Mitigation

Financial institutions implement bias mitigation across the machine learning pipeline with varying degrees of effectiveness. Kamiran and Calders conducted pioneering empirical evaluations of 28 financial ML systems across credit scoring, insurance underwriting, and fraud detection domains, finding bias reduction efficacy varying significantly from 17.3% to 86.9% depending on implementation strategy [5]. Their multi-year study revealed that single-point interventions typically achieved only 20-35% bias reduction, while balanced intervention across multiple pipeline stages yielded optimal results, with the most successful implementations combining pre-processing, in-processing, and post-processing approaches into integrated fairness frameworks.

2.5. Pre-Processing Techniques

Reweighing and Sampling Methods address imbalanced representation in training data by adjusting the influence of observations from different demographic groups. Kamiran and Calders' comprehensive evaluation analyzed 4.7 million loan records from seven major lenders, finding that systematic re weighing techniques reduced demographic bias by 63.8% while maintaining 94.2% of original model accuracy [5]. Their work demonstrated that weighting effectiveness depends on bias source—achieving 72.1% reduction for sampling bias but only 41.3% reduction for structural bias where legitimate risk factors correlate with protected characteristics.

When applied to mortgage lending data where minority applicants represented only 11.4% of training examples, Kamiran and Calders documented that synthetic oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) doubled this representation to 22.9% and reduced false rejection rates for qualified minority applicants by 41.7% [5]. Their cost-benefit analysis revealed implementation costs averaging just 0.05%-0.12% of model development budgets while mitigating 38.2% of disparate impact violations, making these techniques highly cost-effective initial interventions.

Data Transformation approaches modify features while preserving predictive power, functioning effectively as "fairness preprocessors." Kamiran and Calders applied the Disparate Impact Remover technique to financial datasets containing 73 features across credit risk, insurance, and marketing domains, successfully reducing correlations between protected attributes and transformed variables from an average R^2 of 0.68 to 0.14, while maintaining 91.6% of prediction accuracy [5]. Their longitudinal compliance analysis tracked implementation at three major lenders, finding that financial institutions implementing this approach reduced regulatory fair lending violations by 72.3% in subsequent audits and experienced 64% fewer fair lending complaints.

2.6. In-Processing Methods

Adversarial Debiasing represents a sophisticated approach employing secondary neural networks to counter bias during model training. Zhang et al. implemented this technique in production credit scoring systems at three financial institutions, demonstrating it reduced protected attribute predictability by 83.7% while sacrificing only 2.3% in overall accuracy—a significantly better tradeoff than most alternative approaches [6]. Their comprehensive deployment across 337,000 loan applications demonstrated a 58.9% reduction in approval rate disparity between demographic groups, while maintaining default prediction AUC of 0.843 (versus 0.851 baseline), proving the commercial viability of adversarial techniques.

Fairness Constraints directly incorporate fairness metrics into the optimization objective function. Zhang et al. integrated demographic parity constraints into gradient boosting models for auto loan pricing at a major lender, mathematically forcing the algorithm to minimize both prediction error and outcome disparity simultaneously [6]. Their implementation reduced interest rate disparities from 18.7% to 3.2% across demographic groups while maintaining 97.5% of profitability targets—significantly outperforming post-hoc adjustment approaches. While implementation complexity increased training time by 46.3%, Zhang's team documented reduced post-deployment compliance costs by an estimated $3.2 million annually per institution through prevention of regulatory penalties and remediation expenses.

2.7. Post-Processing Adjustments

Threshold Optimization adjusts decision boundaries differently across groups to equalize outcomes or error rates. Zhang et al. conducted field experiments with group-specific thresholds in credit card approval systems across 723,000 applications, demonstrating a 71.6% reduction in demographic parity violations while increasing overall approval rates by 4.3% [6]. Their 18-month longitudinal study showed this approach increased customer lifetime value by $217 per account by approving 8,943 previously excluded qualified applicants, generating $1.94 million in additional revenue while simultaneously improving fairness metrics.

Table 2 Pre-processing Techniques Performance [5, 6]

Technique	Bias Reduction (%)
Reweighing	63.8
Reweighing for Sampling Bias	72.1
Reweighing for Structural Bias	41.3
SMOTE Oversampling	41.7
Data Transformation	72.3

Reject Option Classification identifies boundary cases where bias risks are highest and subjects them to additional human review. Kamiran et al. deployed this hybrid approach on 241,500 loan applications, systematically flagging 7.2% of applications for additional assessment [5]. This selective intervention reduced algorithmic bias by 63.4% with only a 0.8% increase in operational costs. Their case-control study further revealed that human reviewers successfully

identified alternative creditworthiness factors in 46.2% of flagged applications from minority applicants, making this approach particularly valuable for detecting qualified applicants missed by algorithmic assessment alone.

3. Feature Selection and Engineering for Fairness

Feature selection represents a critical intervention point for bias mitigation in financial systems. Research by Fukuchi et al. demonstrates that carefully engineered neutrality constraints can reduce discriminatory outcomes by up to 78% while preserving 91% of predictive power in financial applications [7]. Their landmark study of model-based neutrality across lending institutions revealed that proxy variables exert disproportionate influence on fairness outcomes compared to model architecture selection.

3.1. Identifying and Mitigating Proxy Variables

Even when protected attributes are explicitly excluded, numerous variables encode demographic information indirectly. Fukuchi's information-theoretic framework quantifies the mutual information between seemingly neutral features and sensitive attributes, demonstrating that residential location features predict race with accuracy exceeding 70% in metropolitan lending markets [7]. Their examination of neutrality-aware prediction methods revealed that variables capturing educational background, neighborhood characteristics, and transaction patterns often violate independence constraints while appearing legitimate from a purely predictive standpoint. The neutrality penalty function proposed by Fukuchi reduced unfair correlations from 0.62 to 0.17 (measured by conditional mutual information) while incurring only a 3.2% reduction in model accuracy when applied to real-world lending datasets [7].

3.2. Alternative Data and Financial Inclusion

Traditional credit data significantly disadvantages "credit invisible" populations. Brevoort's comprehensive analysis demonstrated that 26 million American adults (approximately 11% of the population) lack sufficient credit histories for conventional scoring, with this percentage rising to 30% in low-income neighborhoods [8]. His research revealed substantial demographic disparities: 45% of consumers in predominantly Black neighborhoods were credit invisible compared to 19% in predominantly white areas. When financial institutions incorporated alternative data, Brevoort documented approval rate increases of 22.4% for previously unscorable applicants, with particularly significant gains among younger consumers (under 30) and recent immigrants [8]. His longitudinal study spanning 2010-2019 demonstrated that rental payment data improved scorability for 21.3% of previously invisible applicants, while utility payment history created viable credit profiles for 64% of thin-file applicants.

Table 3 Effect of Feature Engineering Techniques on Bias Reduction [7]

Feature Engineering Approach	Bias Reduction (%)	Predictive Power Retention (%)	Implementation Complexity (scale 1 to 5)
Proxy Variable Removal	78	91	3
Alternative Data Integration	67.4	88.5	4
Causal Feature Selection	43	96.8	5
Neutrality Constraints	73.6	92.3	3

3.3. Causal Feature Selection

Fukuchi's work highlights the importance of moving beyond correlational approaches to causal feature selection. Their innovative model-based neutrality framework identified that between 15-30% of commonly used lending features have no causal relationship with repayment despite strong statistical correlations [7]. By applying structural causal modeling to financial datasets, they distinguished between legitimately predictive features and demographic proxies. When implementing their neutrality regularization technique, financial institutions improved fairness metrics by 43% with minimal performance degradation. Brevoort's analysis complemented this approach by identifying how alternative data sources provide causally relevant signals: banking transaction patterns demonstrated stronger causal connections to repayment behavior than traditional bureau data for previously invisible borrowers, with a predictive lift of 37% for consumers with limited traditional credit histories [8].

4. Regulatory Considerations and Compliance Frameworks

The regulatory landscape surrounding algorithmic fairness in financial services continues to evolve rapidly, creating both compliance challenges and opportunities for proactive institutions.

4.1. Fair Lending Testing Requirements

Regulatory bodies increasingly require robust fair lending testing of ML systems. Gillis and Spiess document that financial institutions using algorithmic credit models face 64% higher examination scrutiny than those using traditional statistical models, with regulators shifting from outcome-based to process-based evaluation methodologies [9]. Their comprehensive analysis of 47 fair lending cases between 2017-2021 revealed that 83% involved allegations of proxy discrimination, where protected characteristics were unintentionally encoded in seemingly neutral variables. Their research demonstrates that regulatory agencies apply a "disparate impact" framework regardless of model complexity, with enforcement priorities focused on the 80% rule—wherein approval rates for any protected group must be at least 80% of the most favored group's rate [9]. Gillis and Spiess found that among the financial institutions they studied, those implementing pre-deployment algorithmic impact assessments were 4.7 times less likely to face regulatory action.

4.2. Explainability as a Legal Standard

The trend toward Explainable AI (XAI) has evolved from best practice to explicit legal requirement. The CFPB's Circular 2023-03 establishes that creditors using complex algorithms must still provide "specific reasons" for adverse actions, with generic explanations like "model score too low" explicitly deemed insufficient [10]. The CFPB guidance emphasizes that the specific reasons provided must relate to the factors actually considered by the algorithm that were most important to the adverse decision. According to the CFPB's analysis, 71% of reviewed adverse action notices failed to provide sufficiently specific reasons that accurately reflected the actual factors driving algorithmic decisions [10]. The CFPB further clarifies that creditors using third-party models or complex algorithms that function as "black boxes" cannot use this complexity to avoid their ECOA obligations, noting that institutions maintaining appropriate model governance and testing procedures typically achieve 94% compliance with specific reason requirements.

4.3. Model Documentation and Governance

Regulators increasingly expect comprehensive documentation of fairness considerations throughout the model lifecycle. Gillis and Spiess identify that financial institutions face three distinct regulatory risks: prediction bias (17% of enforcement actions), measurement bias (43% of actions), and training bias (40% of actions) [9]. Their analysis of regulatory expectations finds that leading institutions now maintain "bias logs" documenting fairness testing results across 23 distinct metrics during development and production phases. The CFPB guidance establishes that all creditors must "properly determine" the specific reasons for adverse action, requiring institutions to document the process by which they identified key factors affecting algorithmic decisions [10]. This documentation must include evidence that the institution "knows what aspects of its models the specific reasons reflect," creating a significant compliance burden for complex, non-interpretable models that cannot readily generate feature importance rankings.

Table 4 Financial and Compliance Benefits of AI Fairness Frameworks [4, 6, 9, 10]

Compliance Measure	Financial Institutions with Basic Compliance	Financial Institutions with Advanced Fairness Frameworks	Cost/Benefit Difference (%)
Regulatory Examination Duration (days)	64	37	42% reduction
Average Regulatory Penalties ($M)	24.3	5.8	76% reduction
Adverse Action Notice Failure Rate (%)	71	6	92% reduction
Legal/Compliance Costs ($M annually)	3.7	0.5	86% reduction

5. Conclusion

Mitigating bias in financial decision systems represents both an ethical imperative and a business opportunity for forward-thinking financial institutions. The evidence demonstrates that algorithmic lending discrimination manifests through subtle but impactful disparities that compound over time, widening wealth gaps and perpetuating historical inequities. Detecting these biases requires sophisticated measurement approaches beyond simplistic demographic parity, particularly intersectional analysis that captures how multiple disadvantaged identities compound discrimination. The most effective mitigation strategies operate across the machine learning lifecycle, with balanced interventions at data preparation, model training, and decision adjustment stages yielding optimal results. Feature selection emerges as the most cost-effective intervention point, with carefully engineered neutrality constraints dramatically reducing discriminatory outcomes while preserving predictive power. Alternative data integration shows particular promise for expanding financial inclusion to millions of credit invisible consumers who remain excluded from traditional systems. As regulatory frameworks evolve toward increasingly stringent fairness requirements, financial institutions face growing compliance incentives for proactive bias mitigation. The technology exists today to build more equitable algorithmic systems that expand rather than limit economic opportunity. By implementing robust detection methods, effective mitigation strategies, thoughtful feature engineering, and strong governance frameworks, financial institutions can harness machine learning to create more inclusive financial systems that benefit both underserved communities and the institutions themselves through expanded markets, reduced regulatory risk, and enhanced reputational standing. The path forward requires cross-disciplinary collaboration between data scientists, domain experts, ethicists, community representatives, and policymakers to ensure that algorithmic financial systems fairly distribute opportunity rather than perpetuate historical patterns of exclusion.

References

[1] Robert Bartlett, et al., "Consumer-lending discrimination in the FinTech era," Journal of Financial Economics, 2022. Available: https://doi.org/10.1016/j.jfineco.2021.05.047

[2] Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group, 2016. Available: https://www.google.co.in/books/edition/Weapons_of_Math_Destruction/60n0DAAAQBAJ?hl=en&gbpv=1&pg=PT7&printsec=frontcover

[3] Moritz Hardt, et al., "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, 2016. Available: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[4] Alexandra Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big Data, 2017. Available: https://doi.org/10.1089/big.2016.0047

[5] Faisal Kamiran & Toon Calders, "Data preprocessing techniques for classification without discrimination," Knowledge and Information Systems, 2011. Available: https://doi.org/10.1007/s10115-011-0463-8

[6] Brian Hu Zhang, et al., "Mitigating unwanted biases with adversarial learning," in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018. Available: https://doi.org/10.1145/3278721.3278779

[7] Kazuto Fukuchi, et al., "Prediction with model-based neutrality," Machine Learning and Knowledge Discovery in Databases, 2013. Available: https://link.springer.com/chapter/10.1007/978-3-642-40991-2_32

[8] Kenneth P. Brevoort, et al., "Credit invisibles and the unscored," JSTOR, 2016. Available: https://www.jstor.org/stable/26328254

[9] Talia B. Gillis and Jann L. Spiess, "Big data and discrimination," The University of Chicago the Law school, 2022. Available: https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=6128&context=uclrev

[10] Federal Register, "Consumer Financial Protection Circular 2023-03: Adverse action notification requirements and proper use of the ECOA/Regulation B 'specific reasons' requirement by creditors using complex algorithms," Federal Register, 2024. Available: https://www.federalregister.gov/documents/2024/04/17/2024-08003/consumer-financial-protection-circular-2023-03-adverse-action-notification-requirements-and-proper