Fairness-Aware Credit Scoring
A Comparative Study and a Hybrid Model Approach

NIYATI MALIK* and NIKET PATHAK*, University of Illinois Chicago, USA

The use of machine learning in credit scoring raises concerns about fairness and discrimination, particularly toward protected groups such as younger applicants. This study presents a comparative evaluation of one baseline model and eleven fairness-aware models across four real-world credit datasets. Models are assessed using Accuracy and group fairness metrics.

To support a holistic evaluation, we propose a novel composite metric—Responsibility Score (RS)—which integrates accuracy and fairness into a single interpretable value. Additionally, we introduce a hybrid fairness approach that combines a pre-processing method with an in-processing method to mitigate bias both before and during training. This hybrid pipeline was applied to the HMDA dataset and fine-tuned to achieve an optimal trade-off between fairness and performance.

Our evaluation across the datasets shows that pre-processing models, particularly LFR and Reweighing, consistently achieve strong accuracy and fairness balance. Models like LFR and Equalized Odds ranked highest overall based on Responsibility Score. The hybrid model further improved group and individual-level fairness (e.g., SPD, DI, counterfactual fairness), with minimal loss in predictive accuracy. This work provides a practical framework for ethical model selection in high-stakes credit decision systems and demonstrates the value of combining multiple fairness strategies for enhanced outcomes.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Social and professional topics** → Codes of ethics.

Additional Key Words and Phrases: Fairness-aware machine learning, Credit scoring, Algorithmic bias, Fairness metrics, Disparate impact, Statistical parity, Equal opportunity, Hybrid model, Responsibility Score

ACM Reference Format:

Niyati Malik and Niket Pathak. 2025. Fairness-Aware Credit Scoring A Comparative Study and a Hybrid Model Approach. 1, 1 (May 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Machine learning is increasingly being used to automate credit scoring decisions due to its efficiency and predictive capabilities. However, these models often inherit historical bias present in financial data, which can result in discriminatory outcomes—especially against protected demographic groups such as younger applicants, women, or racial minorities. Despite regulations like the Equal Credit Opportunity Act in the U.S., such bias can persist in algorithmic systems, leading to unequal access to financial opportunities.

To mitigate these issues, fairness-aware machine learning methods have been developed at various stages of the modeling pipeline—pre-processing the data, modifying the learning algorithm itself, or post-processing the model's predictions. However, evaluating these techniques remains a challenge due to the complex trade-off between fairness and accuracy, as well as the use of multiple fairness metrics.

*Both authors contributed equally to this research.

Authors' Contact Information: Niyati Malik, nmali@uic.edu; Niket Pathak, npath5@uic.edu, University of Illinois Chicago, Chicago, Illinois, USA.

Manuscript submitted to ACM

In this study, we evaluate twelve models—including a baseline (XGBoost) and eleven fairness-aware models—across four real-world credit datasets. To simplify comparison, we propose a composite metric called the *Responsibility Score (RS)* that combines accuracy and fairness into a single interpretable value.

We further develop a *hybrid model*, combining two existing models, and apply it to the HMDA dataset.

Our work provides a practical framework for evaluating fairness-aware models in high-stakes domains and contributes to the development of responsible credit scoring systems.

Research Questions. To guide our evaluation and analysis, we address the following research questions:

- **RQ1:** How do fairness-aware ML models perform across multiple real-world credit datasets in terms of both accuracy and fairness?
- **RQ2:** Can a composite metric (Responsibility Score) better capture the balance between fairness and accuracy?
- **RQ3:** Which fairness strategy—pre-, in-, or post-processing—offers the most consistent and responsible outcomes?
- **RQ4:** Can a hybrid model combining multiple fairness strategies improve outcomes further?

2 Literature Review

Fairness in machine learning has emerged as a critical area of research, particularly in high-stakes domains like credit scoring, hiring, and criminal justice. This section surveys work across fairness definitions, intervention strategies, and applications in financial systems.

2.1 Fairness Definitions and Metrics

Initial efforts in fairness research introduced formal definitions to quantify group-based disparities. Dwork et al. (2012) proposed the notion of individual fairness—similar individuals should receive similar outcomes—while Barocas and Selbst (2016) focused on group fairness, emphasizing metrics like Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Disparate Impact (DI). These metrics form the basis for most empirical fairness assessments today.

2.2 Fairness Intervention Strategies

Fairness interventions are generally categorized into three families:

- **Pre-processing methods** such as the Disparate Impact Remover (Feldman et al., 2015) and Learning Fair Representations (Zemel et al., 2013) aim to eliminate bias from training data before model training.
- **In-processing techniques**, including Exponentiated Gradient Reduction and Prejudice Remover (Kamishima et al., 2012), embed fairness constraints directly into the training objective.
- **Post-processing approaches** such as Equalized Odds and Reject Option Classification adjust model predictions to improve fairness after model training (Hardt et al., 2016).

Each class of intervention presents trade-offs in terms of fairness, accuracy, interpretability, and feasibility.

2.3 Fairness in Financial Applications

The use of machine learning for creditworthiness prediction has been scrutinized for reproducing historical biases and violating regulatory fairness (e.g., ECOA, Fair Lending laws). Researchers have evaluated various datasets—such as the

German Credit and HMDA datasets—to detect and mitigate unfairness in model predictions (Louizos et al., 2016; Berk et al., 2017).

Studies have shown that fairness-enhancing techniques may reduce disparities but often at the cost of model performance. Recent work focuses on hybrid models and composite fairness-performance metrics to balance ethical and business considerations (Corbett-Davies & Goel, 2018).

2.4 Interpretability and Individual Fairness

Post-hoc interpretability techniques such as SHAP (Lundberg and Lee, 2017) have been employed to ensure that fairness-aware models remain transparent and understandable. In parallel, counterfactual fairness (Kusner et al., 2017) addresses individual-level consistency by evaluating prediction changes under hypothetical alterations of protected attributes.

3 Datasets

To conduct a comprehensive evaluation of fairness-aware machine learning models, we utilized five publicly available datasets—four for comparative analysis and one for a focused case study on HMDA dataset.

3.1 Comparative Evaluation of Fairness Aware Models

We used four real-world credit-related datasets, each offering unique geographic and demographic characteristics to test model generalizability across contexts:

- **German Credit Dataset**: Sourced from the UCI Machine Learning Repository, this dataset includes personal and financial details of applicants from Germany labeled as good or bad credit risks. Protected attribute used for fairness analysis is age.
- **Give Me Some Credit**: Collected from Kaggle, this U.S.-based dataset contains credit history information aimed at predicting whether a person might miss a loan payment. The protected attribute considered is age.
- **Taiwan Credit Default Dataset**: This dataset, also obtained from the UCI repository, contains credit card billing and repayment records for clients in Taiwan. The protected attribute taken is age.
- **Credit Card Approval Dataset**: This dataset comprises loan application information including income, marital status, and loan approval outcomes. Age was used as the protected attribute.

3.2 Fair ML Interventions on HMDA Loan Approvals

For the second part, we used the 2018 Home Mortgage Disclosure Act (HMDA) dataset, filtered to include mortgage applications from the state of Illinois. This dataset contains detailed records of applicant demographics, financial attributes (e.g., loan amount, credit score, debt-to-income ratio), and the loan outcome. It offers a rich basis for fairness evaluation by including multiple protected attributes like `derived_race`, `derived_ethnicity`, and others. The HMDA dataset was used to develop and assess a hybrid fairness pipeline that integrates pre-processing and in-processing fairness interventions.

4 Technical Contribution

This section presents the approach used to assess and improve fairness in credit-based machine learning models, combining multiple datasets, fairness techniques, and evaluation metrics.

4.1 Comparative Evaluation of Fairness Aware Models

To comprehensively evaluate fairness-aware machine learning models across varied real-world scenarios, we conducted a comparative analysis using four diverse credit-related datasets. Each dataset represents different geographies and demographic distributions, offering a broad landscape for fairness evaluation.

We implemented a total of eleven fairness-enhancing models, spanning the three intervention strategies in fairness-aware machine learning: pre-processing, in-processing, and post-processing. For each dataset, models were trained and tested using consistent pre-processing pipelines, and the protected attribute age was explicitly accounted for during fairness evaluations.

- **Pre-processing methods:** Learning Fair Representations (LFR), Disparate Impact Remover, Reweighing.
- **In-processing methods:** Exponentiated Gradient Reduction, Grid Search Reduction, GerryFair Classifier, Prejudice Remover, Adversarial Debiasing.
- **Post-processing methods:** Equalized Odds, Calibrated Equalized Odds, Reject Option Classification.

Four widely recognized fairness metrics were used to benchmark performance: **Statistical Parity Difference** (SPD), **Equal Opportunity Difference** (EOD), **Disparate Impact** (DI), and **Accuracy**. In addition to these, we introduced and used a custom composite metric — the **Responsibility Score** (RS) — which integrates both fairness and accuracy into a single value. The RS formula ($\text{RS} = \alpha \times \text{Accuracy} - \beta \times |\text{SPD}| - \gamma \times |\text{EOD}|$) was designed with adjustable weights ($\alpha = 1, \beta = 2, \gamma = 2$) to emphasize fairness violations more heavily than accuracy alone.

For comparison, we calculated average performance scores across all datasets and grouped results by the type of fairness metric. This enabled a detailed understanding of which strategies yield the most balanced, stable, and ethical outcomes.

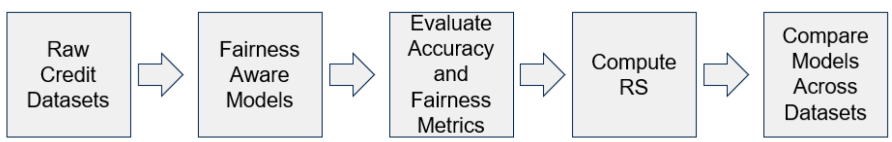


Fig. 1. Basic worflow for Comparative Evaluation of Fairness Aware Models

4.2 Fair ML Interventions on HMDA Loan Approvals

This section focuses on building fair machine learning models using the 2018 HMDA dataset, specifically filtered to include mortgage application records from the state of Illinois. For fairness evaluation, we considered three protected attributes: `derived_race`, `derived_ethnicity`, and `age_favorability`, which were categorized into privileged and unprivileged groups.

We began by developing a baseline machine learning model using a **Random Forest Classifier**, trained on the unprocessed dataset. This model served as a reference point for assessing disparities in predictions across the protected attributes using standard fairness metrics.

To address potential unfairness, we implemented a hybrid fairness pipeline that integrates both pre-processing and in-processing techniques. As a pre-processing step, we applied the **Disparate Impact Remover** (DIR) to reduce bias in the training set by modifying attribute values that are correlated with protected groups. Two levels of intervention were tested: full repair (`repair_level = 1.0`) and partial repair (`repair_level = 0.7`), to allow for flexibility in controlling the trade-off between fairness and data distortion.

Following feature repair, we trained fairness-aware models using **Exponentiated Gradient Reduction** (EGR), an in-processing method that adjusts the training process to account for group disparities. This method wraps around the base classifier and iteratively updates model weights in response to fairness-driven learning objectives. The repaired datasets from the DIR step were used as inputs to train these EGR-enhanced models.

To fine-tune the fairness–accuracy trade-off in the partial repair model, we applied a **Grid Search** over epsilon (ϵ) values within the EGR framework. This allowed us to systematically explore how varying ϵ impacts the model's performance and fairness, ensuring that the partial repair configuration operated at an optimal level rather than relying on a fixed ϵ value.

We developed a third hybrid model in which we performed a **Grid Search** over ϵ values within the **Exponentiated Gradient** framework to identify the configuration that maximized the **Responsibility Score**. This allowed us to fine-tune the model to achieve the best balance between fairness and performance.

Beyond group fairness, we also incorporated counterfactual fairness analysis on partial repair model to assess how model predictions behave under hypothetical changes to protected attributes. This involved generating counterfactual examples for each test instance by altering only the protected attribute while keeping all other inputs constant, and observing whether the model's prediction changed.

In parallel, to support model transparency, we conducted **SHAP** (SHapley Additive Explanations) analysis. SHAP values were calculated to understand the contribution of each input feature to individual predictions. This helped evaluate whether the model relied on financial reasoning or was indirectly influenced by proxy attributes linked to protected characteristics.

Overall, this multi-step methodology integrated fairness interventions, optimization, counterfactual reasoning, and explainability to develop models that are fair, flexible, and interpretable.

Fig. 2. Basic worflow for Fair ML Interventions on HMDA Loan Approvals

5 Results and Evaluation

We present the outcomes of our experiments aimed at understanding how different fairness-aware machine learning models perform in real-world credit scoring scenarios. This includes both a comparative evaluation across four diverse datasets and a targeted fairness optimization pipeline applied to the HMDA 2018 dataset. The results address our core research questions (RQ1–RQ4).

5.1 RQ1: Model Performance Across Datasets

We compared all models using four key metrics: Accuracy, Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Disparate Impact (DI). Our key findings were:

- **LFR** consistently achieved high accuracy while also maintaining low SPD and EOD values, making it the most balanced model overall.
- **Equalized Odds** achieved the best group fairness, particularly minimizing Equal Opportunity Difference, though sometimes with a small trade-off in accuracy.
- **Reject Option Classification** displayed highly unstable behavior, with Disparate Impact values exceeding legal thresholds, raising concerns about practical usability.
- **Grid Search** and **Exponentiated Gradient** also performed well, offering a balanced trade-off between accuracy and fairness metrics.
- **Calibrated Equalized Odds** and **Prejudice Remover** had lower accuracy and inconsistent fairness performance.

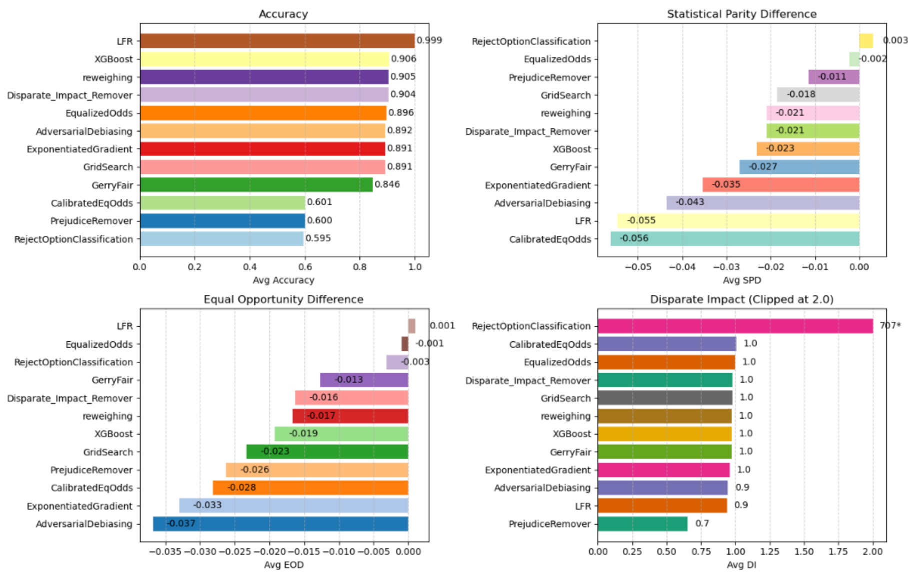


Fig. 3. Average Accuracy, SPD, EOD, and Disparate Impact for all models across four credit datasets. Disparate Impact is clipped at 2.0 for readability.

5.2 RQ2: Composite Metric for Fairness and Accuracy

To address contradictions between fairness metrics, we proposed the Responsibility Score (RS):

$$RS = \alpha \times \text{Accuracy} - \beta \times |\text{SPD}| - \gamma \times |\text{EOD}|$$

Using weights $\alpha = 1$, $\beta = 2$, $\gamma = 2$, RS emphasized fairness violations more than prediction error.

Interpretation:

- RS allowed us to rank models holistically, providing a single interpretable score that reflects accuracy and ethical responsibility.
- Models like LFR and Equalized Odds consistently ranked highest.
- RS penalized highly accurate but biased models.

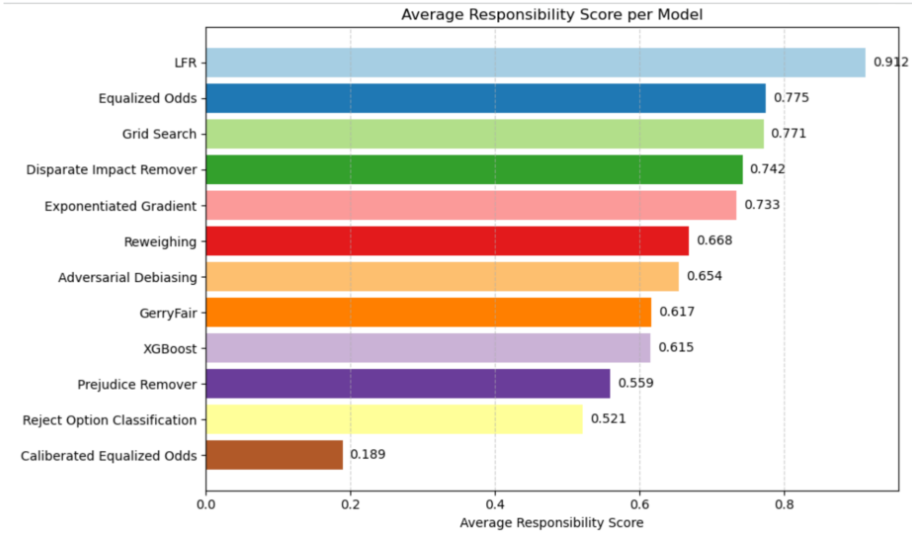
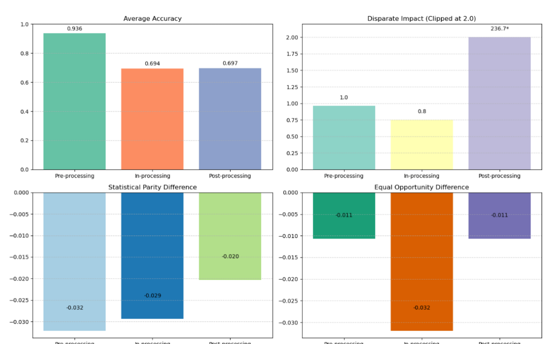


Fig. 4. Average Responsibility Score for all models across four credit datasets

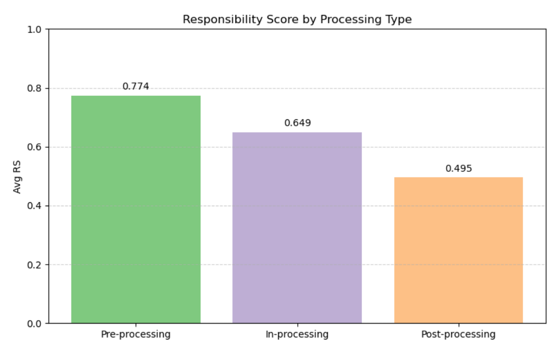
5.3 RQ3: Strategy-wise Comparison

We grouped models by mitigation strategy: pre-, in-, and post-processing.

- **Pre-processing models** had the highest average Responsibility Score (RS) and the least variance across datasets.
- **In-processing models** offered reasonable trade-offs between fairness and accuracy.
- **Post-processing models** were less stable and often reduced accuracy.



(a) Average Accuracy, SPD, EOD, and Disparate Impact for all types of processing models across four credit datasets.



(b) Average Responsibility Score for all types of processing models across four credit datasets.

Fig. 5. Fairness and performance metrics for pre-, in-, and post-processing models across datasets.

5.4 RQ4: Hybrid Model on HMDA

We implemented a hybrid pipeline combining DIR and EGR on the 2018 HMDA dataset from Illinois.

5.4.1 Baseline Model Performance. The Random Forest baseline had 98.31% accuracy but failed fairness metrics. It demonstrated substantial SPD and EOD, especially against `derived_race`.

Class	Precision	Recall	F1-Score	Support
0 (Denied)	0.95	0.97	0.96	15,067
1 (Approved)	0.99	0.99	0.99	54,200

(a) Classification performance of the baseline model on the HMDA dataset.

Protected Attribute	Disparate Impact (DI)	Statistical Parity Difference (SPD)	Equal Opportunity Difference (EOD)	Average Odds Difference (AOD)
derived_race	0.7657	-0.1861	-0.0089	-0.0118
derived_ethnicity	0.8975	-0.0807	-0.0089	0.0023
age_favorability	0.8830	-0.0928	-0.0002	-0.0090

(b) Fairness metrics across three protected attributes for the baseline model.

Fig. 6. Performance and fairness metrics of the baseline model (Random Forest Classifier) on the HMDA dataset.

5.4.2 Full Repair (repair_level = 1.0, ϵ = 0.10). This model achieved near-perfect fairness (DI ≈ 1, SPD ≈ 0), but recall for class 0 dropped to 8%.

Metric	Accuracy	DI	SPD	EOD	AOD
Value	0.7951	0.9822	-0.0174	-0.0007	-0.0053

Model Evaluation Summary

Class	Precision	Recall	F1-Score	Support
0.0	0.78	0.08	0.15	15067
1.0	0.80	0.99	0.88	54200

Classification Report

Fig. 7. Performance and Fairness Metrics for Full Repair Hybrid Model

Interpretation: Full repair overcompensates, erasing useful patterns, leading to overly cautious classifiers.

5.4.3 Partial Repair (repair_level = 0.7, ϵ = 0.05). This version achieved 95.9% accuracy with marked fairness improvements (DI ≈ 0.94, SPD ≈ -0.047).

Interpretation: This model best balanced fairness with performance, demonstrating strong deployment potential.

5.4.4 Responsibility Score Optimization: Third Model. By performing a grid search over ϵ in EGR, we maximized RS directly. This yielded the most ethical model configuration.

Interpretation: Parameter tuning tailored to a composite fairness-performance objective (RS) leads to more informed trade-offs.

Metric	Accuracy	DI	SPD	EOD	AOD
Value	0.9588	0.9385	-0.0474	0.0098	0.1634

Model Evaluation Summary

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.94	0.91	15,067
1.0	0.98	0.96	0.97	54,200

Classification Report

Fig. 8. Performance and Fairness Metrics for Partial Repair Hybrid Model

5.5 Counterfactual Fairness Evaluation

Counterfactual fairness checks whether altering a protected attribute changes the outcome. The WCF Score = 0.9249 indicated high individual fairness.

Race Group	Total Samples	Changed Predictions	Unfairness Rate
0.0 (Unprivileged)	6,478	808	12.47%
1.0 (Privileged)	62,789	4,366	6.95%

Fig. 9. Counterfactual Fairness evaluation for RS Optimized Model.

Interpretation: Model predictions were stable across protected attribute variations, affirming ethical consistency.

5.6 Model Interpretability with SHAP

SHAP analysis showed that variables like `loan_amount`, `dti_clean`, and `property_value` had the highest influence. Protected attributes had negligible impact.

Interpretation: The model learned fair decision boundaries centered around financial logic, not demographic proxies.

These results underscore the importance of combining model-agnostic fairness techniques, composite evaluation scores, and interpretability tools to ensure ethical deployment of credit-based machine learning systems.

6 Technical Challenges

Throughout the project, we encountered several technical challenges that shaped the development of our methodology:

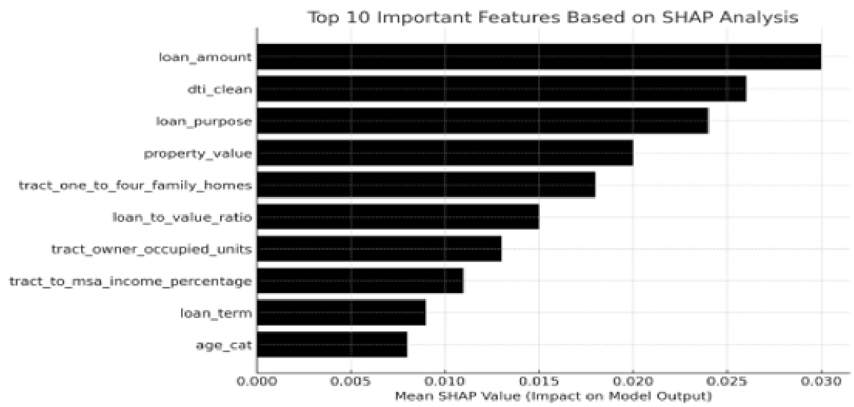


Fig. 10. SHAP Analysis for RS Optimized Model.

- **Metric Contradictions**: Different fairness metrics often provided conflicting evaluations. This made it difficult to compare models holistically, motivating the development of a unified composite measure— Responsibility Score.
- **Heterogeneous Datasets**: The four datasets varied widely in terms of target labels, feature distributions, and protected attributes. This required careful preprocessing, normalization, and attribute mapping to ensure consistent analysis.
- **Ensuring Transparency**: Maintaining interpretability after applying fairness interventions was another challenge. Techniques such as SHAP had to be integrated to analyze the influence of features post-intervention.
- **Resource Constraints**: Due to limited computational resources, training multiple models, running SHAP and counterfactual analyses often took several hours. In many instances, notebooks crashed mid-execution, requiring repetitive reruns and checkpointing to ensure progress.

7 Conclusion

This study examined how fairness-aware machine learning can be effectively integrated into credit decision systems. We evaluated twelve models across four real-world credit datasets and introduced a new composite metric—Responsibility Score (RS)—to jointly assess fairness and predictive accuracy. RS enabled a unified comparison across models, addressing the challenge of conflicting fairness metrics.

For a deeper investigation, we applied a hybrid fairness pipeline on the HMDA dataset by combining Disparate Impact Remover and Exponentiated Gradient Reduction. Among the configurations tested, the partial repair model (repair level = 0.7, $\epsilon = 0.05$) offered the best balance between fairness and accuracy, outperforming both baseline and full repair setups.

To further optimize fairness without sacrificing performance, we fine-tuned model parameters to maximize RS directly. This configuration achieved strong ethical and predictive outcomes. Additionally, we used SHAP and counterfactual fairness analysis to ensure that the model's decisions were transparent and not driven by bias-related proxies.

In conclusion, our work demonstrates that with the right tools and metrics, fairness-aware credit scoring is both practical and impactful. The hybrid approach, paired with responsibility-driven optimization, offers a promising path toward building more trustworthy financial models.

References

[1] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. *Innovations in Theoretical Computer Science*, 214–226.

[2] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C. (2013). Learning fair representations. *International Conference on Machine Learning (ICML)*, 325–333.

[3] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning (ICML)*, 60–69.

[4] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 35–50.

[5] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems (NeurIPS)*.

[6] Fish, B., Kun, J., Lelkes, Á. (2016). A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 144–152.

[7] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

[9] IBM Research. (2019). AI Fairness 360 Toolkit. Retrieved from https://aif360.mybluemix.net

[10] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, K., Miroglio, B., ... Wallach, H. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Retrieved from https://fairlearn.org

[11] Yeh, I.-C., Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480. Dataset: Taiwan Credit Default, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[12] Dua, D., Graff, C. (2017). UCI Machine Learning Repository: German Credit Data. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[13] Kaggle. (2011). Give Me Some Credit Dataset. https://www.kaggle.com/datasets/c/GiveMeSomeCredit

[14] CFPB. (2018). Home Mortgage Disclosure Act (HMDA) Data. https://ffiec.cfpb.gov/data-publication/

[15] Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems (NeurIPS)*.

[16] Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R. (2016). The variational fair autoencoder. *International Conference on Learning Representations (ICLR)*.

[17] Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. (SHAP-like explanations)

[18] Credit Card Approval Dataset. (2020). Retrieved from Kaggle: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

[19] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785