International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I12P115

Volume 72 Issue 12, 128-132, December 2024
© 2024 Seventh Sense Research Group®

Original Article

Fair or Flawed? Assessing AI's Impact on Credit Decisions

Vikas Agarwal

College of Computing, Georgia Institute of Technology, Atlanta, USA.

[1]*Corresponding Author : vagarwal311@gatech.edu*

Received: 01 November 2024 Revised: 29 November 2024 Accepted: 14 December 2024 Published: 31 December 2024

Abstract - *The adoption of AI in financial decision-making, especially credit scoring, has sparked concerns about fairness and bias in outcomes. This study examines how biases in AI models affect protected groups, exploring fairness metrics and mitigation techniques to address these challenges. Using industry datasets, it highlights the trade-off between accuracy and equity, showcasing ways to design fairness-aware systems. The findings emphasize transparency, continuous monitoring, and ethical practices as critical for responsible AI use in banking. By addressing bias, financial institutions can ensure inclusive and unbiased credit decision processes, balancing performance with equity in the rapidly evolving landscape of AI-driven finance.*

Keywords - *Credit score, Fairness, Bias, AI, Machine learning.*

1. Introduction

Credit scoring is one of the cornerstone components of the financial sector. While 1.5 billion people in the world still do not have access to banking services, the remaining group is eligible for lending. The need for a smart credit scoring system is now more crucial than ever, as traditional credit scoring models overlook complex factors like financial behaviour.[1] Artificial Intelligence (AI) is revolutionizing credit scoring and lending processes, making predictions more accurate. Its ability to analyze large datasets, identify patterns, and make predictions has streamlined decision-making and enhanced efficiency.

However, these advantages often come at the cost of fairness and transparency. Studies have revealed that AI models are more prone to biases because of factors like data quality or algorithm processes, each having a disproportionate impact on protected groups.

According to research from the USC Information Sciences Institute, there can be up to 39% biased data depending on the database studies and the type of metrics considered.[1] These biases can exacerbate social inequalities by limiting financial access for marginalized communities. Current literature often focuses on algorithmic sophistication without adequately addressing the real-world implications of biased outcomes.

Moreover, discussions on effective fairness metrics and mitigation strategies remain fragmented and underexplored, leaving critical questions unanswered about the practical implementation of equitable AI systems in finance.

The present study aims to close these gaps by looking at the underlying sources of bias in AI credit scoring algorithms, assessing fairness criteria, and providing viable mitigation techniques.

The goal of the project is to offer a paradigm for creating fairness-aware AI systems that support inclusivity and performance by fusing ethical considerations with real-world case studies. By doing this, the study hopes to promote a more just financial ecosystem and further the field of credit scoring.

2. The Question at Hand

While we explore the discrepancies between fairness and bias in AI-based credit scoring models, the key question at hand is - *How can biases in creditworthiness predictions be detected, quantified, and mitigated effectively?* At the same time, what can be done to balance fairness, predict accuracy, and maintain profitability in automated decision-making systems?

These questions not only challenge the technical limitations of AI but also touch upon critical ethical, regulatory, and societal dimensions. By addressing these concerns, this research explores how to balance equitable financial inclusion and sustainable business practices, ensuring that automated systems remain transparent, accountable, and adaptable to diverse socio-economic contexts.

3. Methods for Fair and Profitable Credit Risk Prediction

To address fairness while maintaining profitability, a structured approach was employed using the *"Default of Credit Card Clients"* dataset from the UCI Machine Learning Repository. Below is a detailed breakdown of the steps and methodologies adopted.

3.1. Data Collection

The dataset comprised 30,000 observations and 25 variables, including demographic attributes such as age and sex, alongside credit history, payment records, and default status. Key considerations included identifying protected attributes to address potential biases: individuals over 40 were classified as unprivileged, aligned with the Age Discrimination in Employment Act of 1967, while male and female classifications were assessed under the Equal Pay Act of 1963 and the Civil Rights Act of 1964.

For this research, we used a feature of each dataset to determine the creditworthiness of customers while avoiding any dependence on protracted attributes like age or sex. New columns were created based on existing features to derive additional insights, and irrelevant or redundant features were excluded to create a clean dataset.

Additionally, efforts were made to balance the dataset, ensuring adequate representation of both privileged and unprivileged groups to enable unbiased model training.

To protect people from intentional or unconscious discrimination and harm, the law prohibits unfair treatment/decisions made based on human characteristics. These characteristics are recognized in the form of ' protected classes'.

This foundational step was crucial in identifying sensitive features that could influence fairness in predictions.

3.2. Analysis Techniques

The analysis undertakes multiple aspects ranging from creditworthiness to establishing a fairness metric and taking measures for bias mitigation. Creditworthiness was calculated by averaging the customer's past payment records.

This score, which ranges from 0 to 100, indicates the likelihood of a customer defaulting. We used it to examine the profit or loss from loan approvals according to various thresholds of creditworthiness. The profits were calculated by comparing actual defaults against predicted outcomes based on creditworthiness.

A profit matrix was established, taking into account different possible outcomes, such as a default predicted by the model when no default actually occurred or vice versa.

Here is how the calculation of creditworthiness can be visualized:

Table 1. Creditworthiness score

Record No.	Default ('Y')	Credit Avg. (Mean)	Creditworthy-ness score
20557	0	0.000000	75.0
16520	0	1.500000	56.0
16096	0	-1.166667	90.0
3230	1	0.000000	75.0
2163	0	0.333333	71.0

Calculating the profit matrix, which assesses the financial results of various model predictions, is a crucial component of the analysis. For instance, if the model forecasts a customer's default but they do not, a revenue-generating opportunity is lost. On the other hand, if the model predicts that a customer will not default, but they do, the model could suffer a loss due to not taking necessary preventive actions. This matrix helps understand how well the model performs in terms of actual business outcomes.

Table 2. Profit matrix

Dataset Default status	Creditworthiness Default status(X >= Creditworthiness)	Profit
Yes	Yes	+10
Yes	No	-5
No	Yes	-3
No	No	0

Our study employed two metrics, Statistical Parity Difference (SPD) and Disparate Impact (DI), to measure fairness after the profit matrix was established.

The SPD calculates the gap between the privileged and underprivileged groups in terms of the percentage of favourable outcomes (such as being granted credit). This is computed by taking the proportion of favourable outcomes for each group and subtracting one from the other. If the result is zero, it indicates that the model is treating both groups equally. A non-zero value suggests that the model is biased towards one group.

For example, if 70% of the privileged group (those under 40) receives favourable outcomes (such as being predicted not to default), but only 60% of the unprivileged group (those 40 and above) receives the same favourable outcome, the SPD would be 10%. This indicates that the privileged group is benefiting more from favourable outcomes.

DI measures how much more likely the unprivileged group is to receive a favourable outcome compared to the privileged group. It is calculated by dividing the proportion of favourable outcomes for the unprivileged group by the proportion for the privileged group. An ideal DI value is 1, which means that both groups are treated equally. If the DI value deviates significantly from 1, it indicates a potential bias in favour of the privileged group. For instance, if 60% of the unprivileged group (aged 40 and above) receives a favourable outcome, and 80% of the privileged group (under 40) receives the same, the DI would be 0.75. This suggests that the unprivileged group is receiving fewer favourable outcomes relative to the privileged group.

The SPD calculates the gap between the privileged and underprivileged groups in terms of the percentage of favourable outcomes (such as being granted credit). Once the fairness metrics (SPD and DI) are calculated, the next step is bias mitigation. If the model shows that one group is disproportionately benefiting from favourable outcomes, adjustments can be made to the thresholds of the creditworthiness score. The threshold is the cutoff value above which a customer is considered not likely to default. By adjusting the threshold for the unprivileged group, the model can be calibrated to ensure that both groups receive similar rates of favourable outcomes.

For example, if the unprivileged group (aged 40 and above) is receiving fewer favourable outcomes, the threshold for this group can be lowered so that more customers in this group are approved for credit, even if their creditworthiness score is lower than that of the privileged group. The goal is to adjust the thresholds in a way that reduces bias while maintaining business profitability. After adjusting the thresholds, the fairness metrics are recalculated. If the changes have been successful in reducing bias, the SPD should approach zero, and the DI should approach 1, indicating that both groups are now receiving similar treatment from the model. The creditworthiness score's thresholds should be fine-tuned to ensure that both fairness and profitability are optimized.

4. Results and Discussion

4.1. Fair in Focus: Understanding the Results

Based on the fairness metrics, the primary outcome was that both the privileged and underprivileged sections were treated *equitably*.

The DI value is within the acceptable range, reinforcing the finding that the system operates without bias toward either group. Here is a graphical representation of the corresponding data:

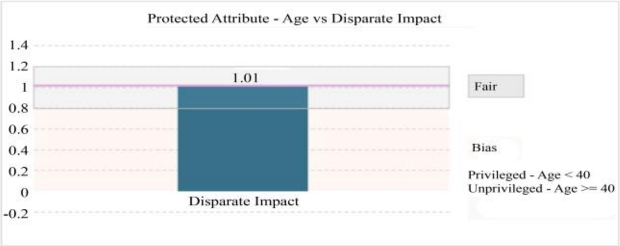

Fig. 1 Disparate impact value

The SPD value also comes close to 0, the acceptable value for fairness. The following diagram on the SPD value further reinforces that both the \ groups receive a fair opportunity with no noticeable bias.

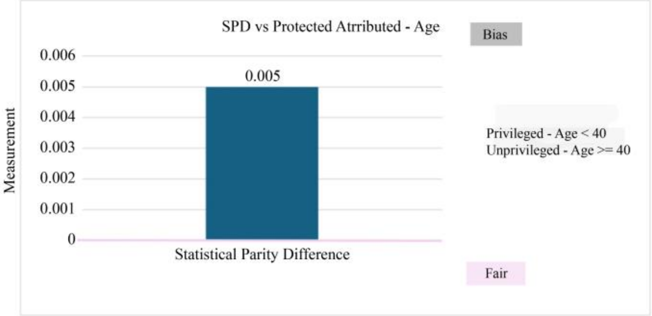

Fig. 2 Statistical parity difference value

4.2. Additional Bias Mitigation

4.2.1. Reweighting

Reweighting entails assigning various weights to instances in the dataset to balance the representation of privileged and unprivileged groups. This method makes sure that the model does not rank results according to innate group differences. To match their influence with that of the privileged group, instances from under-represented groups (such as those 40 years of age and older) were given higher weights in the credit-scoring dataset. This technique successfully addresses unbalances without changing the original data.

4.2.2. Adversarial Debiasing

Adversarial debiasing is using adversarial neural networks to reduce bias during model training. This method uses an adversarial model to identify sensitive attributes (like age or gender) and a primary model to forecast the target variable (like default status). In order to reduce the dependence on sensitive attributes, the primary model is trained to decrease the accuracy of its predictions for the adversarial model. This guarantees excellent predictive accuracy and reduced bias in the credit-scoring projections.

4.2.3. Data Augmentation

Data augmentation techniques can improve the representation of minority groups in the dataset. For instance, it helps resolve class imbalances by creating synthetic data points using methods such as SMOTE (Synthetic Minority Over-sampling Technique). The algorithm is exposed to a more varied dataset by producing artificial examples that reflect marginalized communities, which reduces bias in the results.

4.2.4. Fair Representation Learning

Fair representation learning converts the dataset into a latent space, separating sensitive variables from predictive features. By using this technique, model predictions are guaranteed to be unaffected by protected attributes. This method allowed the credit-scoring model to be fair without compromising its overall accuracy.

4.2.5. Algorithmic Transparency

Implementing transparency tools such as explainable AI (XAI) allows decision-making processes to be inspected for potential bias. Stakeholders can spot and address any unintentional biases in the system by knowing how the model makes certain predictions.

To reduce bias in the model, we set threshold values for both the privileged and unprivileged groups. These thresholds help determine what level of risk or reward each group should receive.

- Privileged Group Threshold: 50
- Unprivileged Group Threshold: 75

We then calculated the maximum profit for each group using a formula. The results were:

- Profit for the Privileged Group: 74,981
- Profit for the Unprivileged Group: 25,810

These results show that the privileged group received much higher profits, which indicates a possible bias towards that group.

These metrics show that the privileged group is benefiting more from the model. The Disparate Impact (DI) for the privileged group is low (0.3969), while the unprivileged group has a higher DI (0.8881), indicating that the unprivileged group is not being treated fairly. Similarly, the Statistical Parity Difference (SPD) is much higher for the privileged group, showing a bias in their favour.

After applying the bias mitigation steps, we observed the following:

The fairness metrics (DI and SPD) no longer show significant bias between the groups, but this does not necessarily mean the bias is gone. Sometimes, applying mitigation techniques can introduce new biases, especially if the data was already fair before applying the changes.

In this case, the mitigation steps did not make the dataset fairer. Instead, it ended up benefiting the privileged group more. This shows that we need to be careful with bias mitigation techniques. Sometimes, they might not work as expected, and they need to be tested and adjusted carefully.

4.3. Navigating the Ethical and Societal Landscape of AI Credit Scoring

While AI cannot be separated from modern, digitalizing and evolving financial activities like credit scoring, it is crucial to understand its ethical and societal impact. Bias in AI can further deep-rooted economic disenfranchisement, perpetuating cycles of poverty and exclusion.

The societal implications extend beyond the immediate financial harm, as stigmatized groups may face long-term consequences, such as limited access to resources and fewer opportunities for upward mobility. As these biased models become more entrenched in financial systems, it is crucial to consider how they shape social norms and public trust in the fairness of automated decision-making processes.

5. Conclusion

The influence of AI extends far beyond technological advancements. It represents a fundamental shift in how financial institutions assess creditworthiness. By integrating machine learning algorithms, predictive analytics, and alternative data sources, AI has enabled a more nuanced, real-time approach to credit risk evaluation. This evolution has not only improved the efficiency of credit scoring but has also paved the way for greater financial inclusion, allowing more individuals to access credit opportunities. Moving from rigid, rule-based systems to dynamic, adaptable models has enhanced the flexibility of credit assessments, transforming how financial institutions manage risk.

Moreover, the emergence of explainable AI has addressed critical concerns related to transparency and fairness. By ensuring that credit models are not only accurate but also understandable, it fosters trust among stakeholders, including consumers and regulators. The real-world implications of these advancements are profound, offering the potential to revolutionize financial ecosystems, improve risk management, and redefine how credit decisions are made.

As we conclude this review, it is essential to recognize that AI in credit scoring is still evolving, with continuous changes influencing industry practices. Future research should focus on deepening our understanding of the ethical considerations surrounding AI models—particularly fairness, transparency, and regulatory compliance. Moreover, exploring disruptive technologies like decentralized finance (DeFi) and the Internet of Things (IoT) presents exciting opportunities for further investigation.

Industry practices must strike a balance between embracing AI advancements for improved credit assessment and maintaining a commitment to ethical standards and consumer protection. Financial institutions need to navigate the tension between innovation and responsible AI usage. Cross-industry collaboration, as seen in emerging trends, could drive further breakthroughs in credit scoring, creating a more holistic approach to data security, regulatory adherence, and consumer trust.

In conclusion, this review highlights the pivotal role AI plays in shaping the future of credit scoring. Its multifaceted impact—ranging from predictive analytics to ethical considerations—sets the stage for a dynamic and transformative era in financial assessment. As the industry continues to embrace AI, a thoughtful, ethical approach is necessary to harness the full potential of these technological advancements. The journey toward more inclusive, transparent, and accurate credit scoring practices is just beginning, with the promise of expanding financial opportunities for a broader range of individuals.

References

[1] Magali Gruet, ‘That’s Just Common Sense’. USC Researchers Find Bias in up to 38.6% of ‘Facts’ used by AI, USC Viterbi School of Engineering, 2022. [Online]. Available: https://viterbischool.usc.edu/news/2022/05/thats-just-common-sense-usc-researchers-find-bias-in-up-to-38-6-of-facts-used-by-ai/

[2] Wilhelmina Afua Addy et al., “AI in Credit Scoring: A Comprehensive Review of Models and Predictive Analytics,” *Global Journal of Engineering and Technology Advances*, vol. 18, no. 2, pp. 11-129, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Pal Dru Koi, How AI is Revolutionizing Credit Scoring: A Glimpse into the Future of Finance, Medium, 2023. [Online]. Available: https://medium.com/technologyone/how-ai-is-revolutionizing-credit-scoring-a-glimpse-into-the-future-of-finance-5188b60597d6

[4] Role of AI in Credit Scoring, Niyogin, 2024. [Online]. Available: https://www.niyogin.com/blogs/role-of-ai-in-credit-scoring

[5] Faisal Kamiran, and Toon Calders, “Data Preprocessing Techniques for Classification without Discrimination,” *Knowledge and Information Systems*, vol. 33, pp. 1-33, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[6] Alessandro Castelnovo, “A Clarification of the Nuances in the Fairness Metric Landscape,” *Scientific Reports*, vol. 12, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rachel K.E. Bellamy, AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, *arxiv*, pp. 1-20, 2018. [CrossRef] [Google Scholar] [Publisher Link]