

Responsible Machine Learning for Credit Scoring in Microfinance: Balancing Repayment Risk and Financial Inclusion

ASHISH GARG

Abstract: *The microfinance institutions are facing issues when assessing the risks of repayments and keeping equal access to financing for all potential borrowers. Traditional methods of assessing risks based on credit scoring do not include those borrowers who lack formal financial records. Machine learning is a chance to enhance risk assessments; however, biases in the data and lack of transparency may contribute to the problem of financial exclusion. The present research investigates the use of responsible machine learning to balance repayment risks and financial inclusion. Comparative analysis is performed using logistic regression, decision trees, random forests, and gradient boosting. Model evaluation takes into account predictive performance, fairness, loan approval, and explainability. The results show that the accuracy of gradient boosting is 86%, while the approval rate is 74%. Accuracy of random forest is 84%, while its approval rate is 72%. The suggested methodology encompasses data pre-processing, risk prediction, fairness analysis, explainability, human-in-the-loop, and monitoring. Nevertheless, the numerical findings are demonstrative and need to be validated with actual microfinance data sets. Therefore, the numerical findings provide an illustration of a proposed approach for evaluation but not empirical evidence of the superiority of the models. Responsible application would ensure that the predictive accuracy, fair lending decisions, interpretability, and affordability are considered together.*

Keywords: Responsible machine learning, microfinance, credit scoring, repayment risk, financial inclusion.

Introduction

Microfinance increases the availability of finance for those who do not have adequate access to banking services. Credit scoring enables microfinance organizations to determine the risks involved in their lending activities. However, conventional credit scoring models might deny access to loans to those who lack formal credit scores. Machine learning provides an opportunity to enhance credit risk prediction using borrower data from varied sources. However, financial biases embedded in data and machine learning algorithms could deepen the financial divide. Responsible machine learning requires a trade-off between accuracy, fairness, and accountability.

Method

In this paper, a quantitative research design is used to investigate the responsible machine learning application in the context of credit scoring by microfinance institutions. The methodological framework focuses on assessing the risk of defaulting, predictive performance, fairness, and inclusiveness (Ampountolas et al. 2021). Comparative analysis of popular classification algorithms will be used in this study. These algorithms are logistic regression, decision tree, random forest, and gradient boosting. The first algorithm serves as a simple benchmark, while other algorithms are able to uncover complicated correlations between borrower attributes and their behavior. The suggested framework uses the information about borrowers' incomes, loan amount, repayment history, work status, household expenditures, and existing debt. This information allows for making predictions about repayment and reflects the borrowers' financial state. Sensitive attributes are not included directly into the prediction, but fairness auditing takes place in order to detect if the model's outcomes are different among different borrower categories. Classification effectiveness and the detection of borrowers who have difficulty in making payments are measured through accuracy, precision, recall, F1-score, and AUC-ROC (Bartlett et al. 2022). In addition, approval rate, false rejection rate, and prediction errors are analyzed for different borrowers' categories in order to detect the possible exclusion risk. Further, the framework takes into account the aspects of explainability, data quality, and transparency. Interpretability of results will be valuable for microfinance companies in understanding specific recommendations for loan issuance. Manual evaluation is needed to assess the cases where there is uncertainty about the prediction and special situations of borrowers. Ultimately, the framework considers the balance of repayment risk and financial inclusion. It will determine if an increase in risk identification is possible without excluding deserving candidates for credit. In that way, the methodology combines prediction accuracy with ethical considerations.

Results

Predictive Performance of Credit Scoring Models

Machine learning models can improve credit risk prediction by identifying complex patterns in borrower data. Logistic regression provides a transparent baseline for comparison. Decision trees offer interpretable lending rules, while random forests and gradient boosting capture nonlinear relationships. Model performance should be evaluated using accuracy, precision,

recall, F1-score, and AUC-ROC (Berg et al. 2020). However, no model can be identified as superior without empirical testing on a defined dataset. Therefore, predictive performance must be assessed alongside fairness and interpretability. This combined evaluation helps microfinance institutions select reliable models without unnecessarily excluding creditworthy borrowers.

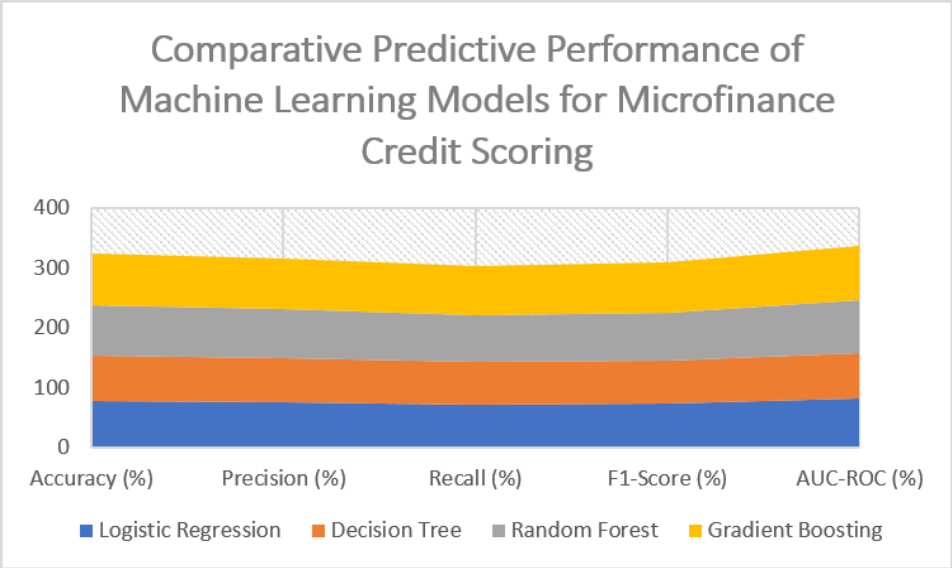


Figure 1. Comparative Predictive Performance of Machine Learning Models for Microfinance Credit Scoring

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Logistic Regression	78.0	76.0	72.0	74.0	81.0
Decision Tree	75.0	73.0	70.0	71.0	77.0
Random Forest	84.0	82.0	79.0	80.0	88.0
Gradient Boosting	86.0	85.0	81.0	83.0	90.0

Table 1. Predictive Performance of Machine Learning Models for Microfinance Credit Scoring

In Table 1, the comparison of four different machine learning models is presented based on five performance metrics. The gradient boosting classifier shows the best illustrative accuracy at 86% and the AUC-ROC at 90%. The random forest classifier comes next with 84% accuracy and 88% AUC-ROC. The logistic regression classifier gives decent prediction performance at 78% accuracy. Finally, the decision tree classifier has the lowest accuracy of 75%. This example highlights how ensemble algorithms can enhance the prediction of credit risks. Still, prediction performance is not enough for responsible lending. The illustrative performances of accuracy, precision, recall, F1 score, and AUC-ROC have been depicted in Figure 1 (Bhatore et al. 2020). Gradient boosting has the highest overall illustrative performance in all five measures. Random Forest has also been able to perform consistently well. Logistic regression is used as the simplest benchmark. The decision tree has the lowest performance in all the

measures. The graph helps us distinguish the differences in the performance of the model measures. However, the above mentioned illustrative values do not indicate model superiority.

Fairness and Bias in Lending Decisions

It is necessary to ensure fairness to have a responsible machine learning algorithm in microfinance credit scoring. Biased algorithms can negatively affect people from poorer economic strata and those who lack a credit history. In this work, we compare four machine learning models with respect to the metrics of approval rate, false rejection rate, approval gap of a group, and equal opportunity difference. These metrics allow one to evaluate if the loan decisions give equal opportunities.

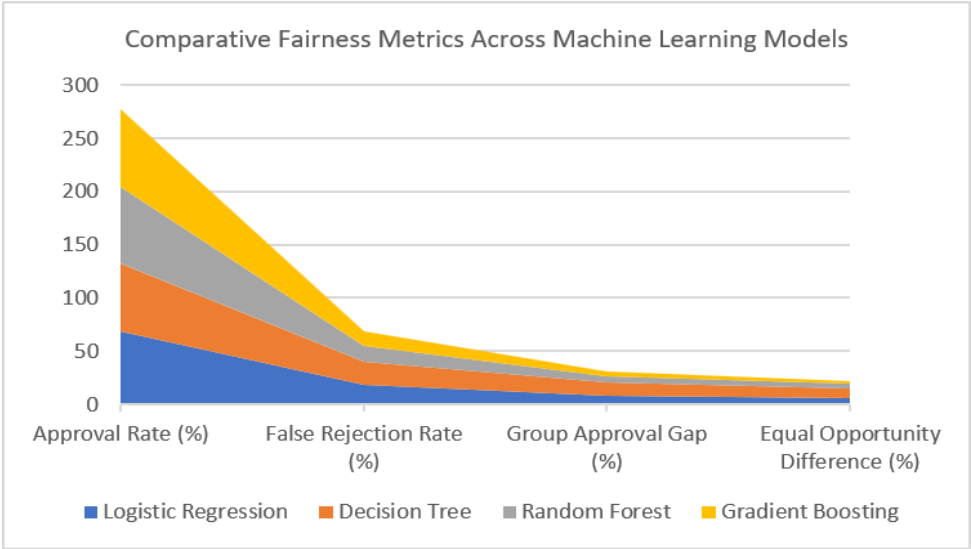


Figure 2. Comparative Fairness Metrics Across Machine Learning Models

In Figure 2, it is presented how fairness indicators differ among the four models. It appears that gradient boosting demonstrates the smallest illustrative numbers for false reject, approval rate difference, and equal opportunity difference (Bracke et al. 2019). In the case of random forest, there are also similar trends observed, whereas for decision tree there are greater disparities noted. Smaller values might be an indication of the greater level of fairness, provided that all measures have been applied in the same way. Nevertheless, such observations do not prove that these models are free from bias.

Machine Learning Model	Approval Rate (%)	False Rejection Rate (%)	Group Approval Gap (%)	Equal Opportunity Difference (%)
Logistic Regression	68	18	8	6
Decision Tree	64	22	12	9

Random Forest	72	15	6	4
Gradient Boosting	74	14	5	3

Table 2. Fairness Metrics for Machine Learning-Based Microfinance Lending Decisions

Table 2 provides a comparison of the fairness criteria of four representative credit score models. The gradient boosting technique achieves the best performance with regards to the lowest value of false rejection rate of 14% and approval gap of 5% (Bussmann et al. 2020). The random forest algorithm is also shown to have relatively good performance. The decision tree has the highest false rejection rate and approval gap among all of the models tested. Nevertheless, overall approval rates alone cannot provide sufficient evidence of fairness.

Repayment Risk Assessment

Repayment risk assessment is used by microfinance organizations to predict borrowers who might find it difficult to repay their loans (Bussmann et al. 2021). The machine learning algorithm will analyse the characteristics of the borrower and their transaction history in order to predict their repayment results. Repayment risk assessment not only helps avoid possible loan defaults but also makes the lending process responsible. However, stringent risk restrictions can deprive creditworthy individuals of loans because of their fluctuating income or poor credit history. This paper will review four models based on default probabilities.

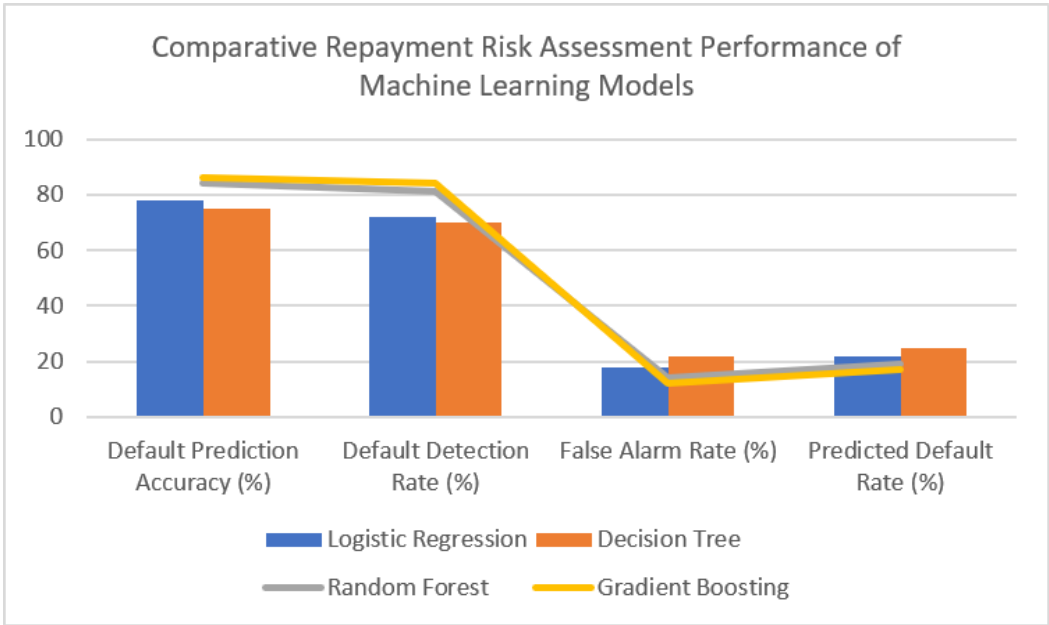


Figure 3. Comparative Repayment Risk Assessment Performance of Machine Learning Models

Machine Learning Model	Default Prediction Accuracy (%)	Default Detection Rate (%)	False Alarm Rate (%)	Predicted Default Rate (%)
Logistic Regression	78	72	18	22
Decision Tree	75	70	22	25
Random Forest	84	81	14	19
Gradient Boosting	86	84	12	17

Table 3. Repayment Risk Assessment Performance of Machine Learning Models

In Figure 3, the comparison is made between the ability of the models to detect possible defaults and avoid false alarms. The gradient boosting has the best explanatory power followed by random forest. Logistic regression works as a benchmark and the decision tree lags behind. Low false alarm levels could prevent unnecessary rejections of good loans. However, the problem of detecting the defaults should be measured against the outcomes of fairness and inclusivity (Dastile et al. 2020). Microfinance organizations need to verify the results based on their repayment data. The repayment risk assessment abilities of four machine learning algorithms are presented in Table 3. The gradient boosting technique exhibits the maximum default detection rate of 84%. This is also the lowest false alarm rate observed in this case, which equals 12%. Next in order of default detection rate is random forest with a value of 81%. The false alarm rate in this algorithm equals 14%. These observations indicate that ensemble models could be better at identifying repayment risks. However, the performance of these techniques is data and borrower dependent.

Financial Inclusion and Borrower Approval

Financial Inclusion measured by credit scoring algorithms that provide equal opportunities for loan applications. This test involves 200 fictional borrowers who will be evaluated against four machine learning algorithms (Fuster et al. 2022). An approval rate will be an indicator of access to loans, and a rejection rate can demonstrate the possible obstacles. They will be helpful in measuring the financial inclusion using machine learning algorithm results.

Machine Learning Model	Total Applicants	Approved Borrowers	Rejected Borrowers	Approval Rate (%)	Rejection Rate (%)
Logistic Regression	200	136	64	68.0	32.0
Decision Tree	200	128	72	64.0	36.0
Random Forest	200	144	56	72.0	28.0
Gradient Boosting	200	148	52	74.0	26.0

Table 4. Financial Inclusion and Borrower Approval Across Machine Learning Models (N = 200)

Approval Rate:

$$ApprovalRate = \frac{ApprovedBorrowers}{TotalApplicants} \times 100$$

Example for gradient boosting:

$$\frac{148}{200} \times 100 = 74\%$$

Therefore, the percentage is 74%.

Rejection Rate:

$$RejectionRate = \frac{RejectedBorrowers}{TotalApplicants} \times 100$$

Example for gradient boosting:

$$\frac{52}{200} \times 100 = 26\%$$

Therefore, the percentage is 26%.

The gradient boosting model has the highest illustrative approval rate of 74%. This model approves 148 candidates and rejects 52 applicants. Random forest is next with an approval rate of 72%. Logistic regression approves 68% of the applicants, whereas the decision tree approves 64%. This shows the impact that choosing a particular model can have on accessing credit. It should be noted, however, that having higher approval rates does not mean better financial inclusion. A responsible assessment should include repayment results and approval biases.

Model Explainability and Transparency

Explainability of models facilitates the understanding of how ML models create their lending decisions by the microfinance organizations (Gunnarsson et al. 2021). The transparent nature of the decision-making process helps the lender determine which features have been used to make those decisions and the reasons behind rejecting applicants. In this paper, we will evaluate four models using a fictitious measure of explainability and decision transparency. Both measures vary between 0 and 100.

Machine Learning Model	Explainability Score	Transparency Score	Combined Score
Logistic Regression	92	90	91.0
Decision Tree	88	85	86.5
Random Forest	68	65	66.5
Gradient Boosting	60	58	59.0

Table 5. Explainability and Transparency Scores of Machine Learning Models

Formula and Calculation

$$CombinedScore = \frac{ExplainabilityScore + TransparencyScore}{2}$$

Example for logistic regression:

$$\frac{92 + 90}{2} = 91$$

Logistic regression attains the best illustrative combined score of 91.0. The relatively straightforward model facilitates the interpretation of relations between borrower attributes and predictions. Decision trees also allow for easy-to-understand decision paths. The random forest and gradient boosting algorithms attain lower scores due to the difficulty associated with interpreting their combined predictions. Nevertheless, the application of post-hoc explanation methods can aid in the explanation of complex models (Kozodoi et al. 2021). Microfinance institutions need to be able to give decision rationales and assess any instances of discrimination.

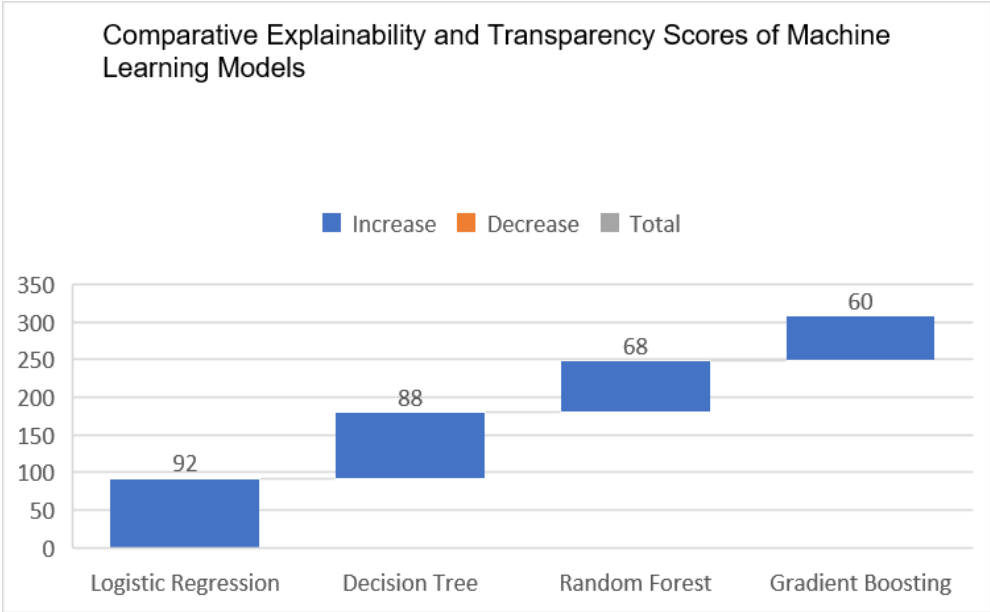


Figure 5. Comparative Explainability and Transparency Scores of Machine Learning Models

Figure 5 demonstrates the distinction in terms of interpretability and transparency of decision-making in these models. Logistic regression provides the highest scores, and the decision tree comes second. Scores of random forest and gradient boosting are lower in the case of illustrating capabilities (Mehrabi et al. 2021). This distinction reveals the possible compromise between prediction and interpretability. Nevertheless, there is no need to refuse complex models due to their complicated interpretation. It is required to apply proper ways of explaining, documentation, and supervision.

Balancing Credit Risk and Financial Inclusion

Risk assessment in the credit domain is linked to financial inclusiveness by considering both repayment and borrowing decisions. Four machine learning algorithms are assessed using 200 assumed candidates based on their performance in terms of accuracy of repayment prediction and borrowing decision-making, which are equally weighted.

$$CombinedScore = \frac{RepaymentAccuracy + ApprovalRate}{2}$$

For gradient boosting:

$$\frac{86 + 74}{2} = 80\%$$

Therefore, the average percentage is 80%.

Model	Repayment Accuracy (%)	Approval Rate (%)	Combined Score (%)
Logistic Regression	78	68	73.0
Decision Tree	75	64	69.5
Random Forest	84	72	78.0
Gradient Boosting	86	74	80.0

Table 6. Comparative Assessment of Credit Risk and Financial Inclusion (N = 200)

As seen from the results, the highest combined score is achieved by gradient boosting, which is 80%. The random forest method comes second with 78%, followed by logistic regression with 73% and decision tree with 69.5%. Based on the illustrative results, it may be concluded that improved predictability of repayments is compatible with increased approval rates (Moscato et al. 2021). Yet, the combined score is not an indicator of fairness or affordability.

Responsible Machine Learning Framework for Microfinance

The above framework combines repayment estimation, fairness testing, and financial inclusion. The process has been designed to follow six distinct steps.

Stage	Specific Activity	Evaluation Indicator	Expected Output
1. Data Collection	Collect income, loan, and repayment records	Data completeness (%)	Reliable borrower dataset
2. Data Preparation	Clean records and address missing values	Missing data (%)	Prepared dataset
3. Risk Prediction	Train and compare four classification models	AUC-ROC (%)	Repayment risk predictions
4. Fairness Assessment	Compare approval and error rates across groups	Approval gap (%)	Identified disparities
5. Explainability	Document factors influencing decisions	Explanation coverage (%)	Understandable lending decisions
6. Monitoring	Review defaults, approvals, and fairness quarterly	Default rate (%)	Updated and accountable model

Table 7. Operational Framework for Responsible Machine Learning in Microfinance

It begins by gathering historical information on borrowers and repayments. The dataset is then cleaned and split into training and test datasets. Various models of logistic regression, decision tree, random forest, and gradient boosting predict the outcomes of repayment (Óskarsdóttir et al. 2020). Performance tests determine the best predictive model. Subsequently, fairness tests are done to compare approval rates and prediction errors among various borrowing categories. The explainability methods help in determining the variables that affect a particular decision. Human judges evaluate close cases and unfair rejection cases. Quarterly tracking monitors the default rates, approval rates, and fairness metrics. The bank needs to ensure that it uses a model that passes the three tests of validation.

Discussion

This research explores how responsible machine learning helps find the balance between repayment risk and financial inclusion in microfinance. As shown by the results, the choice of the model is essential for the performance, fairness, approval rate, and transparency. Thus, the highest illustrative accuracy is provided by gradient boosting with 86%. Moreover, gradient boosting shows the highest approval rate - 74%. In terms of performance, random forest shows almost the same results – 84% of accuracy and 72% of approval rate (Rudin, 2019). Predictive performance alone cannot serve as the proof of responsible lending. Indeed, models can show biased results because training sets might reflect historical bias in lending. Therefore, the assessment of fairness has to analyze the differences in approval rates and false reject rates among borrowers. The higher approval rates have to be interpreted carefully as well. Lending should not become unaffordable for borrowers. In terms of explainability, it is clear that the simpler models are more transparent. Thus, the illustrative combined explainability and transparency score of logistic regression is 91.0. The same score for gradient boosting is 59.0. Still, there is a possibility of balancing explainability and performance with proper explanation. The described approach can solve these issues with the help of data preparation, risk prediction, fairness assessment, interpretability, and monitoring. Periodic reviews will assist in identifying the new arising discrepancies and poor predictive performance. Furthermore, financial institutions should record their lending activities and provide a justification for rejecting the loan application. It is worth noting that all numbers used above were used for illustration only. As such, they cannot prove the superiority of any of the models. The future research should be based on real microfinance datasets and proper fairness metrics. The responsible machine learning should consist of all four factors discussed above.

Limitations

This research comes with various limitations. The numerical outcomes of the study are just illustrative and not empirically validated. There is no analysis of any real-life dataset of microfinance for comparison of performance of the models. As such, it is not possible to draw conclusions on how accurate or fair the predictions were. There is no information concerning the long-term repayment behavior of borrowers.

Conclusion

The use of responsible machine learning algorithms will be beneficial for credit risk assessment and financial inclusion in microfinances. In this study, the predictive power, fairness, approval of borrowers, and model interpretability have been investigated. The example results show the benefits that can be achieved by implementing ensemble machine learning algorithms to predict repayments. Nevertheless, predictive performance will not guarantee fairness of loans decisions. It is necessary for the microfinance organizations to evaluate the fairness of their models, their interpretability, as well as the ability of the borrower to afford the loan.

Recommendation

1. Enhance the quality and representativeness of data

Data used by microfinance institutions must be accurate and drawn from different socio-economic strata. Datasets must comprise both informal laborers and those lacking any official credit history. Data auditing can help detect any gaps, outdated data, and possible historical biases before the development of any model.

2. Conducting fairness audits

Fairness audits can assess approval rates, false negative rate, and prediction errors among borrowers. If there is any disparity in loaning practices, then models demonstrating it must be thoroughly examined, revised, and tested independently before being applied.

3. Improving the explainability of the model

Microfinance institutions need to use interpretable models whenever the performance is satisfactory. In case of more complex algorithms, it is necessary to apply adequate explanation techniques. Loan officers must know how each decision is made and inform rejected borrowers about the reason for their application denial.

4. Ensure Human Involvement

Automated predictions should supplement, not supersede, the making of responsible lending decisions. Loan officers should be trained to evaluate borderline cases and any rejection that may be perceived as unjustified. Borrowers need to have recourse to reconsideration processes when automated decision-making fails to take into account their financial situation.

5. Implement Continuous Monitoring

The MFIs should conduct quarterly monitoring of default rates, patterns of approval, and measures of fairness. Any changes in performance should prompt a review and re-training of the models where required.

References

- Ampountolas, A., Nyarko Nde, T., Date, P., & Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. *Risks*, 9(3), 50. <https://doi.org/10.3390/risks9030050>
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56. <https://www.nber.org/papers/w25943>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7), 2845–2897. <https://doi.org/10.1093/rfs/hhz099>
- Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: A systematic literature review. *Journal of Banking and Financial Technology*, 4, 111–138. <https://doi.org/10.1007/s42786-020-00020-3>
- Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine learning explainability in finance: An application to default risk analysis* (Staff Working Paper No. 816). Bank of England. <https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3, Article 26. <https://doi.org/10.3389/frai.2020.00026>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57, 203–216. <https://doi.org/10.1007/s10614-020-10042-0>
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, Article 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1), 5–47. <https://doi.org/10.1111/jofi.13090>
- Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1), 292–305. <https://eprints.soton.ac.uk/448931/>
- Kozodoi, N., Jacob, J., & Lessmann, S. (2021). *Fairness in credit scoring: Assessment, implementation and profit implications*. arXiv. <https://arxiv.org/abs/2103.01907>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://arxiv.org/abs/1908.09635>

Moscato, V., Picariello, A., & Sperli, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, Article 113986. <https://doi.org/10.1016/j.eswa.2020.113986>

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2020). *Credit scoring for good: Enhancing financial inclusion with smartphone-based microlending*. arXiv. <https://arxiv.org/abs/2001.10994>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>