

An Explainable and Fair Machine Learning Approach for Credit Default Prediction

Mohammed Murtaza Tajammul
Department of Computer Science and Engineering
Muffakham Jah College of Engineering and Technology, Hyderabad, India

Abstract - Credit default prediction plays an important role in helping financial institutions assess the risk of loan applicants. Machine learning models can push prediction accuracy higher, but many of them behave as black boxes and give little insight into how a decision was actually reached. That lack of transparency makes their predictions harder to trust, particularly in settings where fairness matters as much as accuracy.

In this study, we build a credit default prediction system on the Home Credit Default Risk dataset and compare Logistic Regression, a Multi-Layer Perceptron (MLP), XGBoost, and a tuned XGBoost model. The best-performing model is then examined using SHAP and LIME to explain both global feature importance and individual predictions, and fairness is checked across gender groups using the Fairlearn library.

The tuned XGBoost model came out on top with a ROC-AUC of 0.757, and SHAP/LIME gave reasonably clear explanations of what was driving its predictions. The fairness analysis found comparable performance across the two groups on the metrics examined.

Taken together, the results suggest that combining predictive performance with explainability and fairness checks produces a more transparent credit risk system — one that is arguably better suited to real-world lending decisions than accuracy alone would indicate.

Keywords — Credit Default Prediction, XGBoost, Explainable Artificial Intelligence, SHAP, LIME, Fair Machine Learning, Credit Risk Assessment.

1. INTRODUCTION

Predicting whether a borrower will repay a loan is one of the most consequential tasks in the financial sector. Banks and lending institutions depend on credit risk assessment to limit losses and make sound lending decisions. Get it wrong in one direction and the lender loses money to default; get it wrong in the other and creditworthy applicants are turned away. As the volume of financial data keeps growing, machine learning is increasingly stepping in to supplement, or replace, traditional credit scoring rules.

Among the many machine learning techniques available, ensemble methods such as XGBoost have shown strong performance on structured financial data. But high predictive accuracy alone is not enough for real-world adoption. Many of these models are effectively black boxes, which makes it hard for institutions, regulators, and applicants themselves to understand why a particular decision was made. That opacity can erode trust in automated systems, especially in a regulated domain like banking.

Explainable Artificial Intelligence (XAI) addresses this by opening up how a model arrives at its output. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are two of the most widely used tools for this — they surface the factors driving both overall model behaviour and individual predictions, letting decision-makers interpret outputs rather than just trusting a score.

Fairness is the other piece that is easy to overlook. Models trained on historical financial data can pick up patterns that disadvantage particular demographic groups, sometimes without anyone intending it. Checking for this is a basic part of responsible AI development, and it helps establish whether a model's outcomes are actually comparable across groups rather than just assumed to be.

A fair amount of prior work looks at machine learning for credit risk, but most of it focuses on squeezing out better predictive performance. Explainability and fairness tend to get studied on their own, if at all, and rarely inside the same pipeline. That gap is really the motivation for this paper: to see what happens when prediction, explanation, and fairness checking are put together in one workflow rather than treated as three separate projects.

This study develops a credit default prediction system on the Home Credit Default Risk dataset. Four models — Logistic Regression, MLP, XGBoost, and a tuned XGBoost — are trained and compared on standard classification metrics. The best of these is then interpreted with SHAP and LIME, and Fairlearn is used to check fairness across gender groups.

1.1 Objectives of the Study

The primary objectives of this research are:

- To compare the performance of multiple machine learning models for credit default prediction.
- To develop a tuned XGBoost model for predicting loan default risk.
- To enhance model interpretability using SHAP and LIME explainability techniques.
- To evaluate model fairness across gender groups using Fairlearn.
- To analyze the trade-off between predictive performance, interpretability, and fairness in credit risk assessment.

2. RELATED WORK

Machine learning has become a popular approach to credit risk prediction largely because it can capture complex relationships in financial data that simpler models miss. Logistic Regression has remained a mainstay because it's simple and interpretable, but ensemble methods have increasingly overtaken it on large, structured datasets.

Lessmann et al. [6] compared several machine learning algorithms for credit scoring and found that ensemble-based models consistently beat traditional statistical approaches. Their work is a good reminder that model choice should follow predictive performance rather than convention. Chen and Guestrin [2] introduced XGBoost itself, a gradient boosting algorithm built for efficiency and scale that has since become one of the go-to models for structured classification problems.

Accuracy is only part of the story, though, and researchers have leaned more and more into explaining what these models are actually doing. Lundberg and Lee [9] proposed SHAP, which gives a theoretically grounded way of attributing a prediction to each input feature. Ribeiro et al. [10] introduced LIME, which explains individual predictions by fitting a simple, interpretable model locally around them. Between the two, you get both a global picture of model behaviour and a local one for individual decisions.

Fairness has picked up similar attention across finance, healthcare, and recruitment. Models trained on historical data can reproduce the biases baked into that history, so checking for this has become a standard part of responsible AI work. Tools such as Fairlearn [13] give practitioners concrete metrics for spotting performance gaps between demographic groups.

What's less common is combining all three — prediction, explanation, and fairness — inside a single pipeline. Most existing work picks one or two. This study tries to close that gap by putting predictive modelling, SHAP/LIME explainability, and Fairlearn-based fairness evaluation into one experimental workflow.

2.1 Comparative Analysis of Existing Studies

Early research on credit risk prediction focused mostly on improving classification accuracy through statistical methods like Logistic Regression. As machine learning matured, ensemble methods — Random Forest, Gradient Boosting, XGBoost — took over as the higher-performing option on large financial datasets. More recently, attention has shifted toward interpretability, fairness, and responsible AI more broadly.

Even so, relatively few studies check whether their high-performing models can actually explain themselves. That matters a lot in finance, where lending decisions may need to be justified to both customers and regulators. SHAP and LIME have become the default tools for this, but most papers use only one of the two, giving either a global or a local view rather than both.

Fairness shows a similar pattern. A handful of recent studies look at algorithmic bias using metrics like Demographic Parity and Equalized Odds, but fairness evaluation is still missing from a large share of credit risk prediction work. Most papers still emphasize predictive accuracy without checking whether that accuracy holds up consistently across demographic groups.

Table 1 lines up representative studies against the dataset, models, explainability techniques, and fairness methods they used. The pattern is clear: prior work tends to focus on one of these three things at a time, whereas the framework proposed here brings all three together.

Reference	Dataset	ML Models	Explainability	Fairness	Main Contribution
Lessmann et al. [6]	Multiple credit datasets	Logistic Regression, Random Forest, SVM	No	No	Benchmark comparison of credit scoring algorithms
Chen & Guestrin [2]	Structured datasets	XGBoost	No	No	Introduced XGBoost for high-performance prediction
Lundberg & Lee [9]	General ML models	XGBoost	SHAP	No	Global feature-importance explanation
Ribeiro et al. [10]	General ML models	Model-agnostic	LIME	No	Local prediction explanations
Bird et al. [13]	Fairness benchmark datasets	Various ML models	No	Fairlearn	Fairness evaluation framework
This study	Home Credit Default Risk	Logistic Regression, MLP, XGBoost	SHAP + LIME	Fairlearn	Integrated prediction, explainability, and fairness framework

Table 1. Comparison of Representative Studies

2.2 Research Gap and Motivation

Although many studies have explored machine learning techniques for credit risk prediction, most primarily focus on improving predictive performance. Other studies investigate explainability or fairness as separate research problems. Relatively few integrate model comparison, explainability, and fairness evaluation within a single workflow. This limits the practical applicability of many existing approaches, where transparency and fairness are becoming increasingly important alongside predictive accuracy.

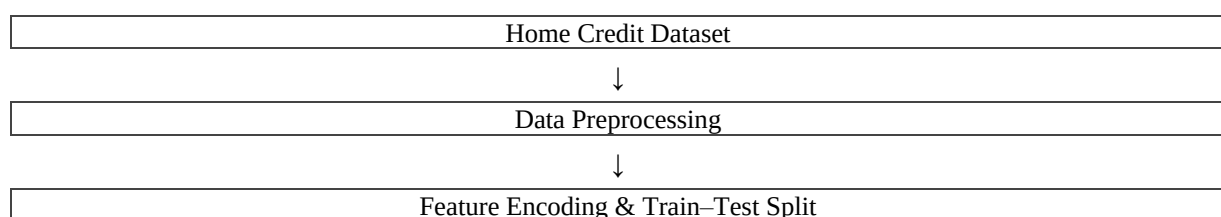
Unlike previous studies that emphasize either prediction accuracy or model interpretability, this work combines predictive modelling, explainability, and fairness evaluation within a single reproducible workflow. Multiple machine learning models are compared to identify the best-performing classifier, after which SHAP and LIME are used to explain model behaviour at both the global and local levels. Finally, Fairlearn is employed to evaluate prediction fairness across gender groups. This integrated approach provides a more transparent and responsible framework for credit default prediction.

Contributions of this study

- Comparative evaluation of Logistic Regression, MLP, XGBoost, and Tuned XGBoost.
- Global explainability using SHAP.
- Local explainability using LIME.
- Fairness assessment using Fairlearn.
- Reproducible implementation using the Home Credit Default Risk dataset.

3. METHODOLOGY

Figure 1 shows the overall workflow used in this study. The process starts with data preprocessing, moves through model training and evaluation, and finishes with explainability and fairness analysis on the best-performing model.



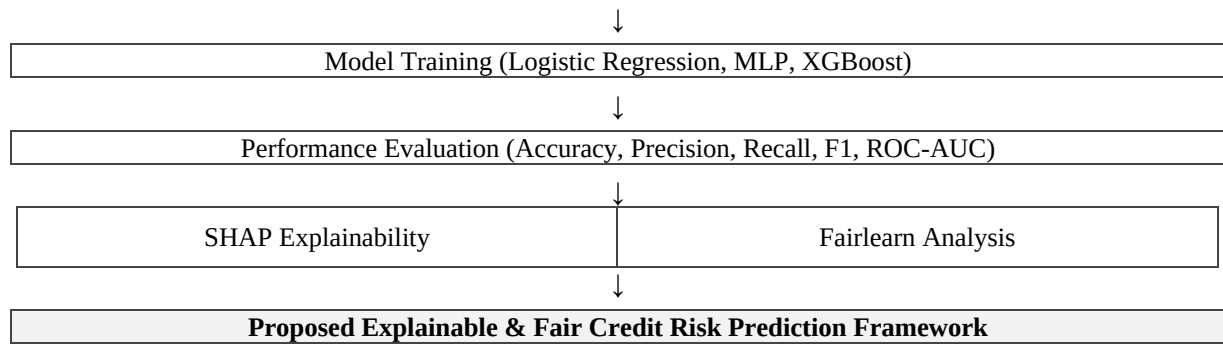


Figure 1. Overall workflow of the proposed explainable and fairness-aware credit default prediction framework.

3.1 Dataset Statistics

Property	Value
Dataset	Home Credit Default Risk
Source	Kaggle
Total Samples (after preprocessing)	61,503
Total Features	172
Target Variable	TARGET
Prediction Task	Binary Classification
Class 0	Non-default
Class 1	Default
Train-Test Split	80:20

Table 2. dataset statistics

Table 2 summarizes the characteristics of the processed dataset used in this study. After preprocessing, the dataset contained **61,503** loan applications with **172** features. The prediction task was formulated as a binary classification problem, where the objective was to identify applicants who were likely to default on their loans.

Although the MLP achieved the highest accuracy, it failed to identify any default cases, resulting in zero precision, recall, and F1-score. This demonstrates that accuracy alone is not an appropriate evaluation metric for imbalanced datasets. The tuned XGBoost model achieved the highest ROC-AUC while maintaining substantially better recall, making it the most suitable model for subsequent explainability and fairness analysis.

3.2 Data Preprocessing

The dataset was first checked for missing values, inconsistent entries, and feature types. Columns with a high proportion of missing values were dropped, and the remaining gaps were filled using appropriate statistical imputation. Categorical variables were one-hot encoded so they could be used by the machine learning algorithms.

The cleaned dataset was then split into training and testing sets so that model performance could be judged on data the model had not seen during training.

3.3 Machine Learning Models

Four classification models were evaluated in this study:

- Logistic Regression
- Multi-Layer Perceptron (MLP)
- XGBoost
- Tuned XGBoost

Logistic Regression served as the baseline for its simplicity and interpretability. The MLP was included to see how a neural network approach stacks up against tree-based methods. XGBoost was chosen for its track record on structured datasets, and it was then hyperparameter-tuned in an attempt to push its performance further.

3.4 Evaluation Metrics

The models were evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Because the dataset is imbalanced — defaults are the minority class — more weight was placed on ROC-AUC and Recall than on Accuracy alone, which can look deceptively good even when a model is essentially ignoring the minority class.

3.5 Fairness Evaluation

To check whether the model behaved consistently across demographic groups, fairness was evaluated using the Fairlearn library [13]. Gender was chosen as the sensitive attribute since it was directly available in the dataset.

Performance was compared between male and female applicants using group-wise accuracy, along with the Demographic Parity Difference and Equalized Odds Difference, to see whether the model's outcomes were comparable across the two groups.

3.6 Hyperparameter Tuning

To improve the performance of the XGBoost classifier, hyperparameter tuning was performed before the final model was selected. Several combinations of model parameters were evaluated, and the configuration that achieved the highest ROC-AUC score on the validation data was selected. The tuned XGBoost model was then used for all explainability and fairness analyses presented in this study.

Hyperparameter	Values
Learning Rate	0.05
Number of Trees (n_estimators)	300
Maximum Depth	6
Subsample	0.80
Column Sample by Tree	0.80
Objective	binary:logistic
Evaluation Metric	auc
Random State	42

Table 3. Tuned XGBoost Hyperparameters

4. EXPERIMENTAL RESULTS

The four models were evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Table 4 summarizes performance on the held-out test set. Figure 2 shows the ROC curve for the proposed model, discussed further below.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.586	0.111	0.591	0.187	0.619
MLP	0.919	0.000	0.000	0.000	0.500
XGBoost	0.919	0.538	0.019	0.036	0.756
Proposed (Tuned) XGBoost	0.717	0.172	0.654	0.272	0.757

Table 4. Performance comparison of evaluated machine learning models

4.1 ROC Curve Analysis

The ROC curves compare how well each model separates defaulters from non-defaulters. The tuned XGBoost model achieved the largest area under the curve of the four, confirming it as the strongest performer on this measure.

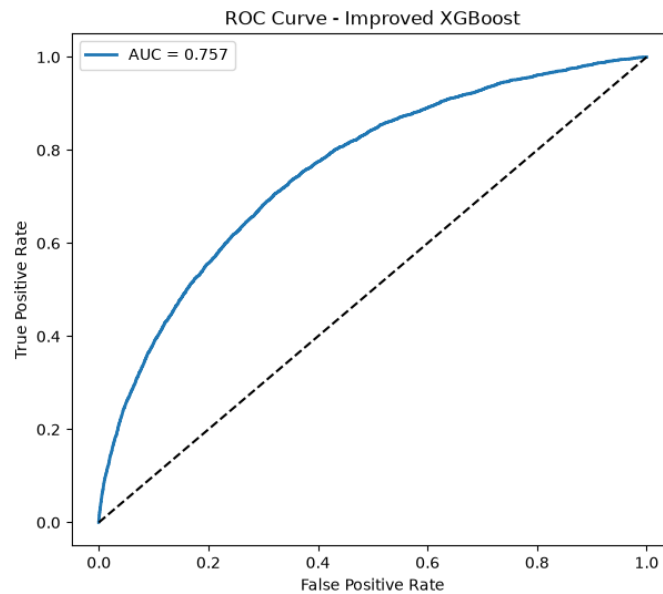


Figure 2. Receiver Operating Characteristic (ROC) curve of the proposed XGBoost model (ROC-AUC = 0.757).

4.2 Precision–Recall Curve Analysis

Because non-default cases considerably outnumber default cases in this dataset, the Precision–Recall curve adds useful information beyond ROC-AUC alone. The tuned XGBoost model held a reasonable balance between precision and recall, catching a meaningful share of default cases without an excessive number of false positives.

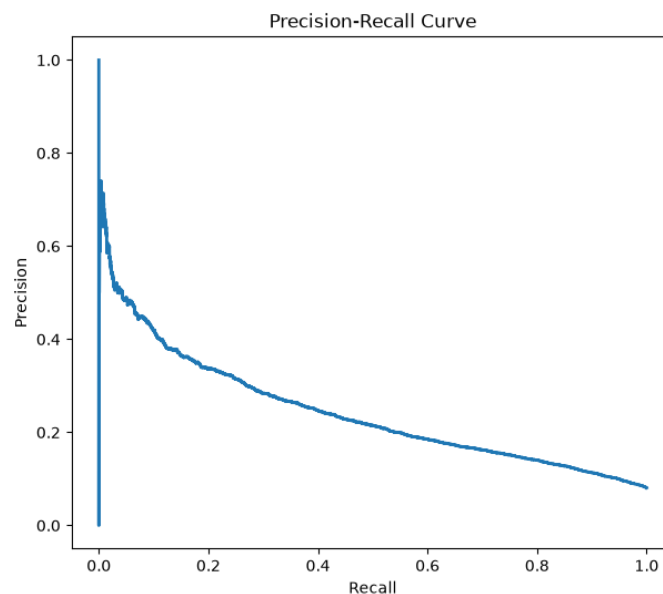


Figure 3. Precision–Recall curve of the proposed XGBoost model.

4.3 SHAP Explainability Analysis

SHAP analysis identified the variables that mattered most to the model's predictions. External credit scores (EXT_SOURCE_3 and EXT_SOURCE_2) and financial variables such as AMT_GOODS_PRICE and AMT_CREDIT stood out as the most influential features — which lines up with the intuitive expectation that prior credit behaviour and financial capacity should drive default risk.

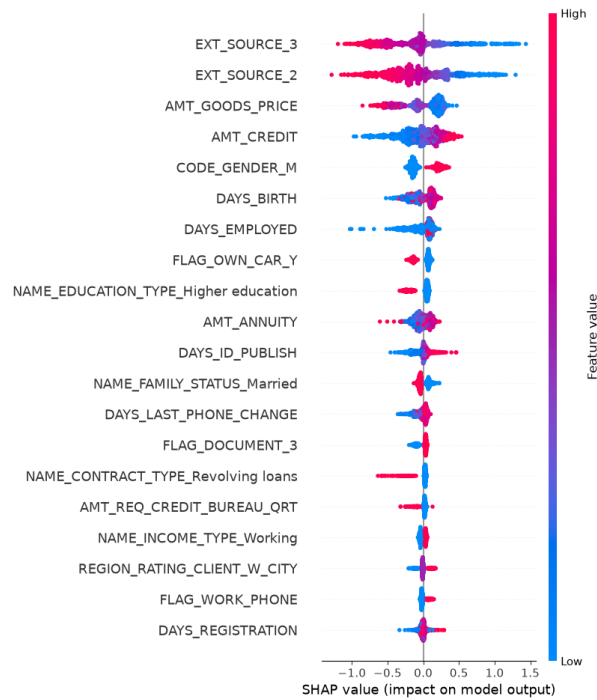


Figure 4. SHAP summary plot illustrating the global feature importance of the proposed XGBoost model.

4.4 LIME Analysis

Where SHAP gives a global view, LIME explains individual predictions. Figure 5 shows how different features influenced the prediction for one applicant. Positive contributions pushed the predicted probability of default up, negative contributions pulled it down — the kind of case-level detail that can help an analyst actually understand a specific decision rather than just trust it.

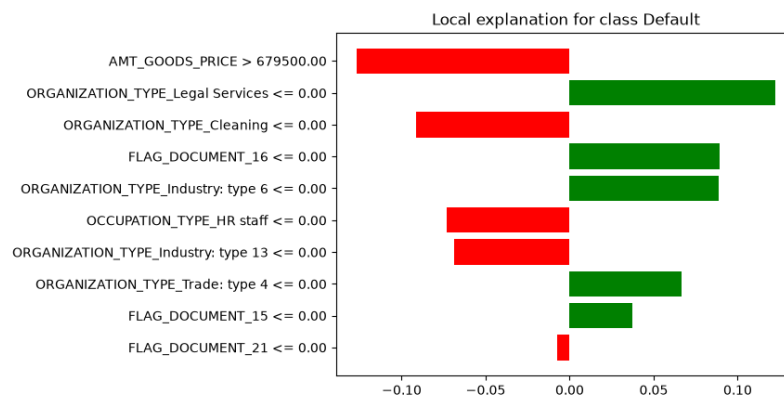


Figure 5. LIME local explanation showing the contribution of individual features to the prediction of a representative loan applicant.

4.5 Fairness Evaluation

The fairness assessment compared prediction performance between male and female applicants using Fairlearn. Both the Demographic Parity Difference and Equalized Odds Difference came out close to zero for the setting examined, which suggests the model was not showing measurable statistical disparity on these particular metrics.

Metric	Value
Female accuracy	0.930
Male accuracy	0.898
Demographic Parity Difference	0.000

Equalized Odds Difference	0.000
---------------------------	-------

Table . Prediction accuracy and fairness metrics across gender groups

Overall, the fairness assessment suggests the framework maintains comparable predictive performance across the two gender groups while meeting the selected fairness metrics at the threshold examined. That said, these results support rather than prove responsible behaviour — the threshold-sensitivity caveat noted here should be checked further in any follow-up validation, and it is worth noting that a small accuracy gap between groups (0.930 vs. 0.898) still exists even though the aggregate fairness metrics land at zero.

5. DISCUSSION

The results make clear that different models behave very differently on an imbalanced credit risk dataset. The MLP reached high accuracy but completely failed to identify the minority default class — a good reminder that accuracy alone can be a misleading metric, and that recall and ROC-AUC matter far more here.

Among the models tested, the tuned XGBoost gave the best overall balance between predictive performance and its ability to actually flag default cases. That's broadly consistent with prior work showing gradient boosting methods do well on structured financial data without needing heavy feature engineering.

The explainability analysis adds a useful second layer on top of the raw numbers. SHAP pointed to external credit scores and financial variables as the biggest drivers of predictions, while LIME showed how those features played out for individual applicants. Together they cover both the global and the local view, which makes the model's behaviour easier to interpret and, hopefully, easier to trust.

The fairness evaluation showed comparable performance across the gender groups on the selected metrics. That's a reassuring result, but fairness outcomes are known to be sensitive to the dataset, the metrics chosen, and the decision threshold used — so these findings should be read as specific to this study rather than as a general claim that the model is fair in every setting.

Limitations

- Only a single publicly available dataset (Home Credit Default Risk) was used, so results may not transfer directly to other lenders or regions.
- Fairness was evaluated with respect to gender only; other sensitive attributes were not examined.
- The problem was framed strictly as binary classification, which does not capture more nuanced risk gradations.
- Fairness metrics were computed at a single decision threshold, and results are known to be threshold-sensitive.
- Results may not generalize to other financial institutions, products, or time periods without further validation.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This study presented an explainable and fairness-aware machine learning framework for credit default prediction on the Home Credit Default Risk dataset. Four models — Logistic Regression, MLP, XGBoost, and a tuned XGBoost — were trained and compared across multiple metrics. The tuned XGBoost classifier came out ahead overall, with the highest ROC-AUC (0.757) and a substantially better recall than the untuned version.

SHAP and LIME were used to make the model's behaviour more transparent, at both the global and local level. The SHAP analysis pointed to external credit scores and financial variables as the strongest drivers of default prediction. The Fairlearn-based fairness evaluation found no measurable statistical disparity across the gender groups examined on the chosen metrics, subject to the threshold-sensitivity caveat discussed in Sections 4.5 and 5.

Put together, this suggests predictive performance, explainability, and fairness can realistically be integrated into a single credit risk assessment system — and that doing so is a step toward more transparent, responsible lending decisions.

6.2 Future Work

- Evaluate the framework on additional public credit datasets to test generalizability across institutions and regions.
- Investigate fairness mitigation techniques (e.g., reweighting, threshold optimization, adversarial debiasing) rather than fairness measurement alone.
- Compare the tuned XGBoost model against CatBoost and LightGBM to see whether the gains here hold up against other gradient boosting implementations.
- Study how fairness metrics shift across different decision thresholds, rather than reporting a single operating point.
- Extend the fairness analysis to additional sensitive attributes such as age, income level, or geographic region.
- Develop a real-time deployment pipeline with automated bias monitoring for production use.

Acknowledgements

The author would like to thank the faculty of the Department of Computer Science and Engineering, Muffakhm Jah College of Engineering and Technology, for their guidance and encouragement throughout this research, and Kaggle for making the Home Credit Default Risk dataset publicly available.

Availability of Code

The complete implementation, trained models, notebooks, and documentation are available at the following GitHub repository:

<https://github.com/mohdmurtaza06/Explainable-Credit-Risk-Prediction>

REFERENCES

- [1] Kaggle, "Home Credit Default Risk," [Online]. Available: <https://www.kaggle.com/competitions/home-credit-default-risk>. Accessed: Jul. 2026.
- [2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- [3] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
- [4] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. Springer, 2021.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
- [6] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.
- [7] A. Bellotti and J. Crook, "Support Vector Machines for Credit Scoring and Discovery of Significant Features," Expert Systems with Applications, vol. 36, no. 2, pp. 3302–3308, 2009.
- [8] D. West, "Neural Network Credit Scoring Models," Computers & Operations Research, vol. 27, no. 11–12, pp. 1131–1152, 2000.
- [9] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, "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, 2016, pp. 1135–1144.
- [11] R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2018.
- [12] B. Hadji Misheva, J. Osterrieder, A. Hirs, O. Kulkarni, and S. F. Lin, "Explainable AI in Credit Risk Management," arXiv:2103.00949, 2021.
- [13] S. Bird et al., "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI," Microsoft, 2020.
- [14] A. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning. MIT Press, 2023.
- [15] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in Credit Scoring: Assessment, Implementation and Profit Implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083–1094, 2022.
- [16] S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," arXiv:2209.09362, 2022.
- [17] Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, "Measuring Fairness in Credit Scoring," SSRN, 2022.
- [18] C. Pérignon and S. Saurin, "The Fairness of Credit Scoring Models," Management Science, 2024.
- [19] R. Bahlool, N. Hewahi, and W. Elmedany, "Performance, Fairness, and Explainability in AI-Based Credit Scoring: A Systematic Literature Review," Journal of Risk and Financial Management, vol. 19, no. 2, 2026.
- [20] V. Sampathkumar, I. P. Pathak, D. K. Ramaraj, R. Kotha, and D. P. Patel, "Federated Explainable AI for Fair and Inclusive Credit Scoring: A Comprehensive Review," IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC), 2026.