Integrating Predictive Accuracy, Fairness, and Explainability in Loan Default Prediction using the Home Credit Dataset

Arvir Jane R. Redondo and Prospero C. Naval Jr.
National Graduate School of Engineering
University of the Philippines Diliman
Email: arredondo@up.edu.ph, pcnaval@up.edu.ph

ABSTRACT

Accurately assessing credit risk is a crucial part of financial decision-making, especially when it comes to determining who qualifies for loans and how to reduce the likelihood of default. In this study, we explored methods for predicting loan defaults using the Home Credit Default Risk dataset available on Kaggle. Our approach focused not only on improving predictive performance but also on addressing fairness concerns across protected demographic groups and ensuring the model's decisions could be explained.

We compared two widely used gradient boosting algorithms, XGBoost and LightGBM, and found that LightGBM offered better predictive results overall. To evaluate fairness, we looked at the False Positive Rate (FPR) gap across gender groups, both before and after applying two fairness-enhancing techniques from the Fairlearn library: the Exponentiated Gradient and Equalized Odds mitigators. These interventions were effective in reducing gender-based disparities with only a minor trade-off in performance.

To understand what influenced the model's decisions, we used SHAP (SHapley Additive exPlanations) values, which highlighted external credit scores, past loan behavior, and borrower age as the most important factors in predicting default. Overall, our findings suggest that machine learning models can be designed to be both accurate and fair, contributing to more transparent and equitable lending practices.

1. INTRODUCTION

Loan default is a significant challenge for credit institutions as it can lead to substantial financial losses if not prevented. To mitigate this risk, lenders conduct thorough assessments to estimate the likelihood of default and make informed decisions about whether to extend credit, often avoiding loans to individuals deemed high-risk. Developing a loan default model involves tackling three key challenges: ensuring predictive accuracy, maintaining fairness across protected groups, and providing explainability to support decision-making and recourse.

Tree-based ensembles like Random Forests, XGBoost, and LightGBM are widely used in credit scoring for their high accuracy and ability to handle class imbalance. Akinjole et al. (2024) report 93.7% accuracy using a stacked model with XGBoost, AdaBoost, and deep learning; Zhu et al. (2024) also utilized LightGBM-XGBoost ensembles. These models capture complex borrower patterns but are harder to interpret, driving interest in explainable AI (XAI).

The Home Credit Default Risk dataset, released on Kaggle as part of a data science competition, has served as a benchmark for evaluating loan default models. It is notable for its rich, relational structure, incorporating multiple tables with historical loan, credit card, and installment payment data. Researchers have applied a variety of ML techniques to this dataset, including feature engineering, dimensionality reduction, and boosting algorithms, to maximize predictive accuracy.

The dataset includes demographic, financial, and behavioral information on applicants. The core application_train table contains the primary target variable TARGET (1 if the loan is in default), and several supplementary tables provide historical data on credit cards, installment payments, bureau credit history, and more. This multi-source format enables the creation of enriched features for predictive modeling.

While many studies have demonstrated strong predictive performance on this dataset, most have focused exclusively on improving accuracy metrics such as ROC AUC, often at the expense of fairness and interpretability. For example, Solanki and Shah (2020) achieved high AUC scores with LightGBM and ensemble models but did not evaluate the fairness of their predictions across demographic groups.

Socioeconomic disparities can affect observed credit scores, but regulators often permit features like credit history even if distributions vary among social groups. Selbst et al. (2019) argue that fairness depends on both the "data frame" (features, labels) and the "sociotechnical frame" (institutions behind the data). For credit, this includes how repayment ability is defined and how unique external situations (e.g. pandemics) distort defaults.

Most research addresses fairness and explainability separately. Some combine two aspects, Pavitha Sugave (2024) integrate "dependency-driven explainable techniques" into a stacked ensemble while Gohar et al. (2022) study fairness in ensemble design. FinRegLab (2021) explores how tools like

SHAP and counterfactuals can support fair underwriting, though they don't propose new algorithms.

A growing body of literature argues for the integration of fairness and explainability into ML models used in high-stakes decision-making, such as lending (Barocas et al., 2019; Mehrabi et al., 2021). Research has shown that algorithmic credit scoring models can inadvertently reproduce or even amplify historical biases present in the data, especially against protected groups such as women or minorities (Hardt et al., 2016). Techniques such as post-processing adjustments (e.g., Equalized Odds) and in-processing methods (e.g., Fairlearn's Exponentiated Gradient) have been proposed to address these issues (Agarwal et al., 2018).

Simultaneously, the need for explainability has prompted researchers to integrate model-agnostic tools such as SHAP (SHapley Additive exPlanations) to provide human-interpretable insights into ML decisions (Lundberg Lee, 2017). These tools are particularly important in finance, where transparency is not only a regulatory requirement but also a factor in user trust.

Despite the availability of these tools, few published studies on the Home Credit dataset have combined all three dimensions, namely predictive accuracy, fairness, and explainability into a single pipeline. This research aims to bridge that gap by building a predictive model, applying fairness constraints with respect to gender, and analyzing feature contributions using SHAP to support interpretable decision-making.

2. OBJECTIVES

This study aims to build a machine learning model for predicting loan defaults using the Home Credit Default Risk dataset from Kaggle, with a focus on three key goals:

- Predictive Accuracy: Develop a high-performing model using leading gradient boosting techniques, specifically XGBoost and LightGBM, by applying feature engineering and hyperparameter tuning to boost predictive accuracy.
- Fairness Across Protected Groups: The study addresses fairness by evaluating whether the model exhibits gender-based bias and applying fairness-aware techniques such as Fairlearn's Exponentiated Gradient and Equalized Odds to reduce any disparities in outcomes
- Explainability: To make the model's decisions more understandable and trustworthy, SHAP values are used to explain predictions, with plans to explore counterfactual explanations using DiCE in future work.

Through these objectives, the study aims to demonstrate that it is possible to create a credit risk model that not only performs well but is also fair and interpretable, addressing key ethical and practical concerns in real-world financial applications.

3. PREPROCESSING

3.1 Table Aggregation

To prepare the dataset, all relevant auxiliary tables (e.g., installments_payments, bureau, previous_application) were aggregated and joined into the application_train table using SK_ID_CURR as the common key. Aggregations included statistical summaries such as mean, sum, and count.

3.2 Feature Engineering

Special emphasis was placed on installment payment behavior, yielding several domain-relevant features:

- PAID_COMPLETE: Total amount of fully paid installments.
- PAID_ONTIME: Count of payments made on or before the due date.
- PAID_COMPLETE_WEIGHTED: Time-weighted sum of complete payments (recent payments weighted higher).
- PAID_ONTIME_WEIGHTED: Time-weighted count of on-time payments.

For the previous applications table, this new feature was also created:

- APP_CREDIT_RATIO: Ratio of the amount requested to the amount granted, representing the optimism of the borrower.

3.3 Encoding, Balancing, and Scaling

- One-hot encoding was applied to categorical features to avoid implicit ordinal assumptions.
- Undersampling of the majority class was employed to balance class distribution.
- Standard Scaling ensured numerical stability across models sensitive to feature magnitude.

3.4 Data Splitting and Feature Selection

A stratified train-test split (80/20) preserved class ratios across subsets. Feature selection was done by removing features with high pairwise correlation, retaining those with the highest SHAP importance scores to maximize interpretability and relevance.

4. METHODOLOGY

4.1 Model Tuning

Two gradient boosting frameworks were evaluated:

- XGBoost: Hyperparameters were optimized using Grid Search (e.g., max_depth, learning_rate, subsample).
- LightGBM: Hyperparameters were tuned with Optuna, an efficient Bayesian optimization tool.

The best model was selected based on cross-validated ROC AUC and F1 scores on the validation set.

4.2 Fairness Enhancement

To address gender-based bias, fairness post-processing techniques were applied using the Fairlearn library:

- ExponentiatedGradient: Reduces demographic disparity by modifying the learning process.
- EqualizedOdds: Adjusts model outputs to equalize false positive/negative rates across gender.

Fairness was assessed pre-and post-intervention, focusing on disparity in False Positive Rate (FPR) between male and female applicants.

5. METRICS

5.1 Model Performance

Performance was measured using the following metrics on the stratified 20% validation set:

- ROC AUC: Measures the model's ability to discriminate between classes.
- Accuracy: Overall correctness of predictions.
- Precision: Correct positive predictions over all positive predictions.
- Recall: True positive rate.
- F1 Score: Harmonic mean of precision and recall.

5.2 Fairness Metrics

The False Positive Rate (FPR) Gap Between Genders measures the absolute difference in the false positive rates for male and female applicants. It reflects how much more likely one gender is to be incorrectly classified as a defaulter. A larger gap indicates potential unfair treatment.

An FPR gap of 5 percentage points or less is considered acceptably fair. This aligns with fairness analysis standards used in prior work, such as Hardt et al. (2016), and reflects a pragmatic tolerance for small disparities.

6. RESULTS AND DISCUSSION

Metric	i	ii	iii	iv	v
ROC AUC	0.6916	0.6916	0.7667	-	-
Accuracy	0.6916	0.6916	**0.7010**	**0.6969**	**0.8057**
Precision	0.6915	0.6915	0.6988	0.6959	-
Recall	0.6920	0.6920	0.7065	0.6993	-
F1	0.6918	0.6918	0.7027	0.6976	-
FPR Gap	NA	NA	**0.1137**	**0.0205**	-

- NA: Not measured
- -: Not available
- i: XGBoost - no feature selection
- ii: XGBoost - w/ feature selection
- iii: LightGBM - w/ feature selection
- iv: LightGBM - w/ feature selection and Exponentiated Gradient
- v: Kaggle 1st place winner

6.1 Performance

The baseline model was built using XGBoost. This initial model achieved a ROC AUC of 0.6916 on the validation set. The performance metrics show that the model struggled to capture the complex patterns in the data, even after balancing the dataset. After performing feature selection by measuring collinearity and SHAP importance, no improvement can be seen from the scores. Since the performance did not worsen either, feature selection is helpful in lessening computational cost due to the high dimensionality of the dataset. However, feature selection was not enough in improving the performance which means that domain knowledge assisted feature engineering is really needed to get high predictive scores.

A comparison between the tuned versions of XGBoost and LightGBM revealed that LightGBM outperformed XGBoost across all major performance metrics including accuracy, precision, recall, F1 score, and ROC AUC. For instance, LightGBM achieved a ROC AUC of 0.7667, compared to 0.6916 for XGBoost.

6.2 Fairness

Before applying any fairness constraints, the LightGBM model showed a significant disparity in False Positive Rate (FPR) between male and female applicants. The FPR gap for Male and Female was 0.1137, indicating bias in assigning False Positives towards Males.

After applying the Exponentiated Gradient fairness mitigation algorithm from the Fairlearn library, the FPR gap narrowed considerably. The FPR gap after applying fairness mitigation was 0.0205 which is below the minimum acceptable FPR gap. While there was a slight reduction in overall performance (e.g., a small drop in Accuracy by 0.0041), this trade-off was considered acceptable in the pursuit of a more equitable model.

6.3 Most Important Features According to SHAP

SHAP analysis provided insights into the model's decision-making process. The most important features included:

Higher values of these features are drivers towards class 0 prediction

- EXT_SOURCE_2: Credit Score from External Source
- EXT_SOURCE_3: Credit Score from External Source
- EXT_SOURCE_1: Credit Score from External Source
- sum_PAID_COMPLETE_WEIGHTED: Total Loans that are fully paid
- AMT_GOODS_PRICE: The price with the goods purchased with the loan

Lower value of these features are drivers towards class 1 prediction

- AMT_ANNUITY: Monthly Payment Commitment

Lower value of these features are drivers towards class 0 prediction

- CODE_GENDER_M: Gender is Male
- OWN_CAR_AGE: Age of the car owned
- PREV_CNT_PAYMENT_MEAN: Average Number of Installment Terms for Previous Loans
- DAYS_BIRTH: Age of Applicant

7. CONCLUSION

This study explored the integration of predictive accuracy, fairness, and explainability in the task of loan default prediction using the Home Credit dataset. The comparative results between XGBoost and LightGBM highlights LightGBM's suitability for structured tabular data and its efficiency in handling large, complex datasets such as this one.

However, the results also suggest that further improvements in predictive performance could be achieved through additional feature engineering. The richness of the Home Credit dataset, particularly the temporal and behavioral data in its auxiliary tables offers more opportunities for extracting informative features that better capture borrower risk.

Importantly, our findings show that it is possible to build an accurate model that remains fair across protected groups, such as gender, when fairness is explicitly considered in the model design process. Techniques like Equalized Odds and Exponentiated Gradient proved effective in reducing disparities without significantly compromising performance.

Finally, the analysis of feature importance revealed that external credit scores, prior loan behavior, and age were among the most influential predictors of default risk. These insights not only validate domain expectations but also support explainability by allowing practitioners to understand and justify the model's decisions.

Overall, this work demonstrates the feasibility of building fair, interpretable, and high-performing models in credit risk prediction, emphasizing that responsible AI in finance is both necessary and achievable.

8. FUTURE DIRECTION

While this study successfully demonstrated the integration of fairness and explainability in credit risk prediction, there remain important opportunities for further research and enhancement.

A particularly promising direction involves the deeper integration of fairness with explainability. While this study utilized SHAP values to interpret the most influential features in the LightGBM model, future work should aim to provide transparent explanations of how fairness interventions such as the Exponentiated Gradient algorithm achieve bias mitigation. Understanding not only how the model predicts, but also how fairness constraints influence decision boundaries, would significantly enhance trust and accountability in high-stakes applications like credit scoring.

Additionally, future work should explore the use of counterfactual explanations for recourse, particularly through tools like DiCE (Diverse Counterfactual Explanations). Unlike global feature importance methods such as SHAP, DiCE allows for individualized, actionable feedback by suggesting minimal changes a rejected applicant could make to receive a favorable decision. This approach not only enhances explainability but also empowers users by offering transparent and fair paths toward credit eligibility.

By combining proactive fairness and personalized explainability, future systems can move closer to the ideal of responsible, user-centered AI in financial decision-making.

9. REFERENCES

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H. (2018). A reductionist approach to fair classification. Proceedings of the 35th International Conference on Machine Learning (ICML), 60, 60–69.

Barocas, S., Hardt, M., Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

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.

FinRegLab. (2022). Explainability fairness in machine learning for credit underwriting: Policy analysis.

Hand, D. J., Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.

Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.

Kazemi, A., Rezvani, M., Shahraki, A. (2021). Predicting loan default using machine learning techniques on Home Credit dataset. International Journal of Advanced Computer Science and Applications, 12(6), 515–523.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Kiradoo, G., Sharma, A. (2024). A comparative analysis of machine learning techniques for loan default prediction. International Journal of Advanced Computer Science and Applications (IJACSA), 15(2), 309–317.

Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136.

Lin, W., Van Biesebroeck, J., Rojas, C. (2024). Explainable machine learning in credit scoring [Preprint]. arXiv. https://arxiv.org/abs/2402.17979

Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1–35.

Salimi, B., Gebru, T., Venkatasubramanian, S., Gummadi, K. P. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. ResearchGate. https://www.researchgate.net/publication/372959063

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT), 59–68. https://www.researchgate.net/publication/330264946

Sharma, M., Verma, N., Nigam, A. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. ResearchGate.

Solanki, V., Shah, R. (2020). Machine learning techniques for Home Credit default risk prediction. International Journal of Advanced Research in Computer Science, 11(3), 1–7.