Quantitative Risk and Fairness in Modern Lending:
An End-to-End Machine Learning Framework Audited Under FEAT Principles

Independent Research Project: Quantitative AI & FinTech

Rajdip Paul (Independent Student Researcher, Class 11; First Author)

Biographical Statement: Rajdip Paul is passionate about both finance and social work. Having competed in investment related competitions at Wharton, he also utilises his knowledge by tutoring disadvantaged children at an NGO. As a past club cricketer, Rajdip applies the same high standards and enthusiasm to all of his aims.

OVERVIEW

This work explores the application of machine learning to credit scoring. A number of different measures are used to evaluate a model's predictive accuracy as well as its fairness and explainability. Models were trained and analysed on a range of credit data-sets sourced from the open literature (Home Credit Default Risk, Give Me Some Credit, Statlog German Credit and Taiwan Default of Credit Card Clients). Model analysis includes an investigation into its predictive performance and fairness, alongside methods to enhance its explainability (using SHAP). Further, group-aware validation and stress testing are explored to determine how robust models are. The developed methodology is then analysed using the four FEAT principles to identify both strengths and gaps in current understanding which so far require more investigation.

SUMMARY

This study develops and evaluates an explainable credit-risk research workflow using the Home Credit Default Risk dataset and three separately modelled public credit-risk datasets: Give Me Some Credit (GMSC), Statlog German Credit, and Taiwan Default of Credit Card Clients. The Home Credit experiment combines applicant-level feature aggregation, an XGBoost classifier, SHAP feature attribution, subgroup fairness audits, sigmoid probability calibration, and controlled feature-perturbation stress tests. The three additional datasets are analysed independently and are not treated as direct external validation of the Home Credit model because their feature spaces, target definitions, populations, and observation windows differ.

The uncalibrated Home Credit model achieved a validation ROC-AUC of 0.7760 and PR-AUC of 0.2742. A sigmoid-calibrated model achieved ROC-AUC 0.7753 and Brier score 0.0668 on the same validation partition. At a threshold of 0.5, the independent experiments achieved ROC-

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

AUC values of 0.8701 for GMSC, 0.7794 for German Credit, and 0.7791 for Taiwan. A duplicate-group sensitivity analysis was conducted for GMSC and Taiwan. Group-aware splits contained no overlapping exact feature groups between training and validation; GMSC ROC-AUC changed from 0.8701 to 0.8623, while Taiwan ROC-AUC changed from 0.7791 to 0.7896. These comparisons indicate broadly similar performance under the tested split strategies, but do not rule out other forms of leakage.

Calibration diagnostics indicated substantial average probability overprediction in the three independent experiments. Subgroup audits identified differences in selection rates, true-positive rates, and false-positive rates; these are descriptive findings and do not establish causal discrimination. A four-principle FEAT self-assessment classified fairness, ethics, accountability, and transparency as partially addressed. The study presents an exploratory research prototype, not a production-ready or regulator-certified lending system. A separate Home Credit base-feature experiment, using only the 120 predictors in application_train.csv and a new stratified 70:15:15 split, was evaluated on an untouched test partition. It achieved ROC-AUC 0.7582, PR-AUC 0.2429, F1 0.3067, and Brier score 0.0679; this is a distinct experiment and not a test of the earlier 194-column engineered model.

Keywords: Credit Risk; Machine Learning; Fairness; Explainable AI; Model Validation.

1. INTRODUCTION

Credit-risk assessment influences access to financial services. Traditional bureau-based systems may provide limited coverage for applicants with thin or incomplete credit histories. Transactional and behavioural information may provide additional predictive signals, but their use raises concerns about data quality, privacy, representativeness, explainability, fairness, and model governance.

Machine-learning methods, including gradient-boosted decision trees, can model nonlinear relationships across heterogeneous borrower attributes. However, predictive performance alone is insufficient for high-impact decision settings. Probability estimates should be assessed for calibration, subgroup outcomes should be examined, and explanations should be interpreted carefully. Validation design is also important: repeated or identical feature records can appear in both training and validation partitions under a row-level random split. Prior work on explainability

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

in financial default modelling has examined input influence and shown that explanations can vary across groups of loans [12].

This project investigates an end-to-end analytical workflow rather than an automated lending product. It combines feature engineering, XGBoost classification, probability calibration, SHAP-based interpretation, subgroup evaluation, controlled stress perturbations, and a project-level assessment informed by the Monetary Authority of Singapore's FEAT principles. Home Credit is treated as the primary experiment. GMSC, German Credit, and Taiwan are trained and evaluated as independent dataset experiments. Their results are not pooled, and no claim is made that one fitted model transfers unchanged across datasets [3].

PaySim was not used as a credit-default validation dataset because its isFraud target represents transaction fraud rather than loan default. LendingClub was also not included in the reported experiments.

2. RESULTS

2.1 Home Credit predictive performance

At threshold 0.5, the confusion matrix was [[41,062, 15,476], [1,560, 3,405]], with rows representing actual non-default/default and columns representing predicted non-default/default. A threshold sweep from 0.1 to 0.7 found the highest F1 among the tested thresholds at 0.70: precision 0.2880, recall 0.3708, and F1 0.3242. The threshold was selected using the validation set and is not an independently tested operating point.

1Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen, “Machine Learning Explainability in Finance: An Application to Default Risk Analysis,” Bank of England Staff Working Paper no. 816 (2019), https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

2Anupam Datta, Shayak Sen, and Yair Zick, “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems,” in 2016 IEEE Symposium on Security and Privacy (2016), 598–617, https://doi.org/10.1109/SP.2016.42.

3Monetary Authority of Singapore, “Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector” (2018), https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat.

2.2 Home Credit explainability, fairness, calibration, and stress tests

The largest mean absolute SHAP values were associated with EXT_SOURCE_2 (0.339), EXT_SOURCE_3 (0.312), EXT_SOURCE_1 (0.137), AMT_GOODS_PRICE (0.121), AMT_CREDIT (0.107), POS_FUTURE_INSTALLMENT_MEAN (0.106), and AMT_ANNUITY (0.101). These values identify influential predictors in the fitted model, not causal drivers of default.

At threshold 0.70, the Home Credit gender audit reported a selection rate of 0.0803, TPR 0.3191, and FPR 0.0624 for the female group (n=40,561). The male group (n=20,940) had selection rate 0.1497, TPR 0.4396, and FPR 0.1168. The XNA group contained only two observations and was unsuitable for stable interpretation. In the age audit, selection rates ranged from 0.2341 for applicants under 25 to 0.0331 for applicants aged 55+. TPR ranged from 0.5883 for applicants aged 25–34 to 0.1736 for applicants aged 55+. These differences warrant contextual investigation but do not prove discrimination.

Before calibration, the validation actual default rate was 0.08073 while the mean predicted probability was 0.39601; the Brier score was 0.18441 and ECE was 0.31528. The sigmoid-calibrated model achieved ROC-AUC 0.77532, Brier score 0.06681, ECE 0.00288, and mean predicted probability 0.07962 on the original validation set. The validation partition had already been used for threshold analysis and earlier diagnostics, so this is not a final untouched test estimate. ECE depends on binning and does not prove universal calibration.

In the preliminary sensitivity tests, baseline mean predicted PD was 0.07962. Reducing income by 20% increased mean predicted PD to 0.08015; increasing credit amount by 20% increased it to 0.08224; applying both changes increased it to 0.08278. The tests did not adjust every correlated feature—for example, annuity and goods price remained unchanged—and should not be described as fully realistic borrower or macroeconomic scenarios.

2.3 Independent dataset model results

GMSC produced the highest ROC-AUC (0.8701) and balanced accuracy (0.7916) among these independent experiments, but its precision was 0.2176. German Credit produced the highest F1 (0.5882) and precision (0.4839), but its validation set contained only 200 observations, making estimates comparatively uncertain. Taiwan produced ROC-AUC 0.7791 and F1 0.5391. Differences in prevalence, sample size, labels, and feature spaces mean these results should not be interpreted as a definitive ranking of model quality.

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

2.4 Comparative calibration

All three independent models overpredicted average default probability on their validation sets. GMSC showed the largest ECE (0.25363), followed by Taiwan (0.20217) and German Credit (0.15900). These values use ten equal-width probability bins and are bin-dependent. The external experiments did not undergo a separate held-out calibration procedure; their Brier and ECE results indicate that probability calibration requires further work.

2.5 Comparative fairness audit

The GMSC age audit showed a selection rate of 0.4507 and FPR 0.3918 for ages 25–34, compared with a selection rate of 0.1083 and FPR 0.0897 for ages 55+. The under-25 group contained 427 observations, so its estimates should be interpreted cautiously. In German Credit, combined personal_status_sex categories A91 and A94 contained only 9 and 12 validation observations, respectively; their rates are too unstable for strong conclusions. In Taiwan, the male group (n=2,402) had selection rate 0.3226 and FPR 0.2281, compared with 0.2685 and 0.1744 for the female group (n=3,598). Taiwan age-group metrics also varied, including a selection rate of 0.3863 and FPR 0.2698 for under-25 applicants.

These audits are descriptive, use a fixed 0.5 threshold for the three independent models, and do not establish causal discrimination, legal non-compliance, or the appropriateness of a lending policy. Small subgroup counts, differing label definitions, proxy variables, and the absence of uncertainty intervals limit interpretation.

2.6 Duplicate-group validation sensitivity

The GMSC group-aware rerun reduced ROC-AUC by 0.00782 and F1 by approximately 0.00467 relative to the original row-level split. The Taiwan group-aware rerun increased ROC-AUC by 0.01048 and F1 by 0.01218. Both group-aware splits had zero exact feature-group overlap between training and validation. The results were therefore broadly similar across the two split strategies in these experiments. This finding does not prove that the original estimates were free from all leakage or dependence effects: the audit addressed exact repeated feature rows only, and the group-aware split used one selected fold rather than repeated grouped cross-validation.

2.7 Separate Home Credit base-feature untouched test experiment

This additional experiment used application_train.csv without the auxiliary-table aggregates used in the primary engineered experiment. The untouched test partition contained 46,127 applicants and had an observed default rate of 8.073%. At the validation-selected threshold of

0.15, the model achieved ROC-AUC 0.7582, PR-AUC 0.2429, precision 0.2289, recall 0.4643, F1 0.3067, and balanced accuracy 0.6635. The Brier score was 0.0679; mean predicted default probability was 0.0811 compared with an observed rate of 0.0807. The ten-bin equal-width ECE was 0.00473 (only five bins contained observations). The confusion matrix was [[36,580, 5,823], [1,995, 1,729]], with rows representing actual non-default/default and columns representing predicted non-default/default.

At the same validation-selected threshold, the test-set gender audit showed selection rates of 0.1219 for females (n=30,305; TPR 0.3878; FPR 0.1023) and 0.2438 for males (n=15,821; TPR 0.5617; FPR 0.2071). Age-group selection rates ranged from 0.0550 for applicants aged 55+ to 0.3634 for applicants under 25. These differences are descriptive and do not establish causal discrimination. Small subgroup effects, proxy variables, and the chosen threshold may influence the results.

This result must not be interpreted as the untouched test performance of the earlier Home Credit model. The earlier model used 194 applicant-level columns, including auxiliary-table aggregates, whereas this separate experiment used only the 120 predictors available in application_train.csv. The two experiments therefore have different feature spaces and training partitions.

2.8 FEAT SELF-ASSESSMENT

This is a project-level self-assessment informed by the FEAT principles. It is not a certification, legal opinion, or claim of regulatory compliance. The prototype does not establish compliance with MAS FEAT, Basel standards, or any other regulatory framework [4].

3. DISCUSSION

The experiments illustrate that discrimination, classification performance, and probability calibration are distinct properties. GMSC achieved the highest ROC-AUC in its independent experiment but had relatively low precision and substantial average probability overprediction. German Credit showed stronger precision and F1 at the fixed threshold, but its small validation sample makes estimates less stable. Home Credit demonstrated that sigmoid calibration can materially improve Brier score while leaving ranking performance broadly similar; however,

4Monetary Authority of Singapore, “Principles to Promote Fairness, Ethics, Accountability and Transparency.”

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

reuse of the validation partition for several analyses means a new untouched test partition is needed for final performance claims.

The duplicate-group sensitivity analysis provides additional evidence about validation robustness for GMSC and Taiwan. Removing exact feature-group overlap from the split did not produce a large change in the reported metrics: GMSC performance decreased modestly, while Taiwan performance increased modestly. The results do not support a claim that duplicate overlap alone explains the original performance. Nevertheless, exact-match grouping does not detect near-duplicates, temporal leakage, target leakage, or other forms of dependence. Repeated grouped cross-validation and further data-lineage checks would provide stronger evidence.

Fairness metrics varied across age, sex, and combined personal-status categories. Such differences may reflect outcome prevalence, data collection, historical lending patterns, model behaviour, or feature proxies. The audits identify areas for further investigation but do not determine the causes. The FEAT review therefore remains partial and should be accompanied by governance, human review, and documented limitations.

The additional base-feature Home Credit experiment provides an independent test estimate for a newly trained model using application-level predictors only. Its ROC-AUC (0.7582) is lower than the earlier engineered model's validation ROC-AUC (0.7760), but this is not a controlled comparison: the feature spaces, split seeds, training sizes, and evaluation partitions differ. The base-feature test result must therefore be reported as a separate experiment rather than evidence of performance change caused solely by removing auxiliary features.

3.1 LIMITATIONS AND FUTURE WORK

The study uses public benchmark datasets with different populations, feature schemas, target definitions, and observation windows. The external datasets are independent experiments, not a harmonised external test of the Home Credit model. The original 194-column Home Credit engineered model has no untouched final test evaluation because its validation partition was reused for threshold analysis, calibration diagnostics, fairness audits, and reporting. A separate base-feature experiment was evaluated on an untouched test partition, but it does not validate the earlier engineered model. The group-aware GMSC and Taiwan results are based on one selected fold per dataset and exact feature hashing; they do not rule out all forms of leakage. German Credit has only 1,000 observations, limiting statistical precision. The UCI dataset documentation also supplies a cost matrix that assigns a higher cost to classifying a bad-credit

applicant as good than the reverse; this study reports conventional threshold-based metrics and does not implement that dataset-specific cost matrix, so its German Credit results should not be interpreted as cost-optimised lending decisions. The reproducibility archives preserve predictions, metrics, split information, and selected model/preprocessing artifacts for the separate base-feature untouched-test experiment. However, serialized estimator and preprocessing compatibility is environment-dependent: at least some archived preprocessing objects produced a scikit-learn compatibility error in the audit environment. Consequently, the archived outputs support result inspection and partial reproduction, but do not establish that every pipeline can be re-executed unchanged without restoring the original software environment and source data.

The external models show substantial calibration error and require calibration on separate development data followed by evaluation on untouched test data. Fairness reports should include uncertainty intervals, intersectional analysis where feasible, proxy review, and sensitivity to threshold choices. Stress tests should perturb coherent groups of correlated variables and should not be presented as macroeconomic stress scenarios without a defensible economic design. Further work should include repeated grouped or temporal cross-validation, baseline models, confidence intervals, calibration curves, decision-cost analysis, robust explainability checks, independent review of data licensing and provenance, and reproducible model monitoring. PaySim should only be used for transaction-fraud or behavioural research unless a valid mapping to a credit outcome is separately established. LendingClub remains unused in the reported experiments.

3.2 CONCLUSION

This study reports a Home Credit applicant-level modelling workflow and three separate public-dataset credit-risk experiments. It combines feature engineering, XGBoost modelling, SHAP-based interpretation, calibration diagnostics, controlled stress perturbations, subgroup fairness audits, and a structured FEAT self-assessment. The group-aware sensitivity analysis found no exact feature-group overlap across training and validation for GMSC and Taiwan, while performance remained broadly similar to the original random-split results. The experiments also show substantial calibration problems in the independent datasets and subgroup metric differences that merit further investigation.

The study does not establish causal fairness, production readiness, regulatory compliance, or transportability of one model across datasets. The results should be treated as an exploratory

research prototype and a basis for more rigorous validation, not as a system suitable for real lending decisions. A separate base-feature Home Credit experiment was evaluated on an untouched test set, but the earlier 194-column engineered model still requires its own independent test evaluation.

Additional Home Credit base-feature test results were generated from the archived application_train.csv using a stratified 70:15:15 split (random_state=2026). The test partition was not used for preprocessing fit, calibration, or threshold selection. Reproducibility outputs include the split manifest, test predictions, metrics, subgroup audit, fitted preprocessor, and calibrated model.

4. MATERIALS & METHODS

4.1 Home Credit Default Risk

The primary experiment used application_train.csv, containing 307,511 applicants, together with the available auxiliary tables: bureau, bureau_balance, previous_application, POS_CASH_balance, installments_payments, and credit_card_balance. Auxiliary records were aggregated to applicant level and joined using SK_ID_CURR. The resulting clean applicant-level table contained 307,511 rows and 194 columns, including the target and identifier. The predictor matrix contained 192 columns after removing TARGET and SK_ID_CURR [5].

A stratified 80:20 split with random_state=42 produced 246,008 training and 61,503 validation applicants. The predictors comprised 176 numerical and 16 categorical columns before encoding. Numerical values were median-imputed; categorical values were most-frequent-imputed and one-hot encoded with unknown categories ignored. The preprocessing pipeline was fitted on the training partition only. The transformed training and validation matrices each contained 316 features.

4.2 Give Me Some Credit

The GMSC experiment used cs-training.csv. The identifier column was removed and one invalid age record (age 1) was excluded, leaving 149,999 records. The target was SeriousDlqin2yrs. MonthlyIncome and NumberOfDependents contained missing values and were median-imputed. The original stratified row-level split contained 119,999 training and 30,000 validation records. Duplicate review identified 609 duplicate rows and 37 feature groups with conflicting

5Home Credit Group, “Home Credit Default Risk” [data set], Kaggle competition (2018), https://www.kaggle.com/competitions/home-credit-default-risk.

target labels. Records were retained rather than arbitrarily deleting observations with conflicting labels [6].

4.3 Statlog German Credit

The German Credit dataset contained 1,000 records, 20 predictors, and one class label. Class 2 (bad credit) was mapped to target 1 and class 1 (good credit) to target 0. A stratified 80:20 split produced 800 training and 200 validation records. Seven numerical and thirteen categorical variables were processed using training-fitted imputation, scaling, and one-hot encoding, producing 61 processed features [7].

4.4 Taiwan Default of Credit Card Clients

The Taiwan Excel dataset contained 30,000 observations after the actual column names in the second row were promoted to headers and the ID field was removed. The target default payment next month was renamed TARGET. The default rate was 22.12%, and no missing values were present. The original stratified 80:20 split produced 24,000 training and 6,000 validation records. All 23 predictors were numerical. The saved Taiwan preprocessing artifact records median imputation but does not record a StandardScaler; therefore, standardization cannot be independently verified from the saved artifact alone. Duplicate review identified 35 exact duplicate rows, 108 repeated feature rows across 52 groups, and 21 groups with conflicting target labels. These records were retained [8].

4.5 Duplicate-group sensitivity design

For GMSC and Taiwan, exact duplicate feature rows were grouped using hashes computed from predictor columns, excluding the target. StratifiedGroupKFold with five folds, shuffling, and random_state=42 generated candidate group-aware splits. The selected fold was the candidate with the smallest combined deviation from a 20% validation fraction and the full-dataset default rate. This selection rule was applied to the candidate folds and is reported as part of the

---

6Credit Fusion and Will Cukierski, “Give Me Some Credit” [competition], Kaggle (2011), https://www.kaggle.com/competitions/GiveMeSomeCredit.

7Hans Hofmann, “Statlog (German Credit Data)” [data set], UCI Machine Learning Repository (1994), https://doi.org/10.24432/C5NC77.

8I.-C. Yeh, “Default of Credit Card Clients” [data set], UCI Machine Learning Repository (2009), https://doi.org/10.24432/C55S3H.

validation design. Identical feature groups were therefore assigned wholly to either training or validation. Conflicting target labels within a feature group were preserved.

The selected GMSC group-aware split contained 120,017 training and 29,982 validation records, with default rates of 6.688% and 6.667%, respectively, and zero overlapping feature groups. The Taiwan group-aware split contained 23,997 training and 6,003 validation records, with default rates of 22.074% and 22.306%, respectively, and zero overlapping feature groups. These are sensitivity analyses on the same datasets, not independent external validation.

In addition, a separate Home Credit base-feature experiment was conducted directly from application_train.csv. TARGET and SK_ID_CURR were excluded from the predictors, leaving 120 raw predictors (104 numerical and 16 categorical). The data were stratified into 70% training (215,257), 15% validation (46,127), and 15% untouched test (46,127), using random_state=2026. This experiment did not use the auxiliary-table aggregates and is reported separately from the 194-column engineered Home Credit experiment.

4.6 Modelling and Evaluation Methods

XGBoost binary classifiers were fitted separately for Home Credit, GMSC, German Credit, and Taiwan [9]. Class weighting was used to address class imbalance. Preprocessing was fitted on each training partition and applied to its corresponding validation partition. Unless otherwise stated, threshold-dependent metrics used a probability threshold of 0.5. Evaluation included ROC-AUC, average precision (PR-AUC), precision, recall, F1-score, balanced accuracy, Brier score, and confusion matrices.

The primary Home Credit model used 200 estimators, maximum depth 4, learning rate 0.1, subsampling 0.8, and column sampling 0.8. For sigmoid calibration, the original 246,008-row training partition was divided into a model-fitting subset of 196,806 rows and a calibration subset of 49,202 rows; the XGBoost model was fitted on the former and the sigmoid calibrator on the latter. The validation partition was then used for calibration diagnostics and threshold analysis, so it is not an untouched final test set. The external models used dataset-specific configurations. GMSC used 200 estimators, depth 4, and learning rate 0.05; German Credit used 150 estimators, depth 2, and learning rate 0.03 with additional regularization; Taiwan used 250 estimators, depth 4, and learning rate 0.04 with additional regularization.

9Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York: ACM, 2016), 785–794, https://doi.org/10.1145/2939672.2939785.

For the GMSC and Taiwan group-aware reruns, the original dataset-specific XGBoost configurations were retained. The saved GMSC group-aware preprocessing artifact records median imputation fitted on the new training partition, but it does not record standard scaling; therefore, scaling cannot be independently confirmed from that artifact alone. The Taiwan predictors were numeric, and the same model configuration was fitted on the group-aware training partition.

SHAP TreeExplainer was applied to a 1,000-row Home Credit validation sample [10]. Mean absolute SHAP values describe the magnitude of feature contributions under the fitted model and explainer configuration; they do not establish causal effects. Stress tests perturbed selected input variables and measured changes in predicted risk. They are controlled model-sensitivity exercises, not macroeconomic forecasts.

Calibration was assessed using the Brier score, mean predicted probability, observed event rate, and expected calibration error (ECE). ECE depends on the binning procedure. Fairness audits compared subgroup selection rates, true-positive rates (TPR), false-positive rates (FPR), and precision at the stated threshold. These descriptive comparisons do not identify causal mechanisms.

For the separate base-feature experiment, median imputation was applied to numerical variables and most-frequent imputation plus one-hot encoding to categorical variables. The preprocessing pipeline was fitted only on the training partition, producing 244 encoded features. An XGBoost classifier was trained on the training partition. Sigmoid calibration and threshold selection were performed using validation data only; the threshold maximising F1 over 0.10–0.70 was 0.15. The test partition was accessed only after these choices were fixed.

REFERENCES

Bracke, Philippe, Anupam Datta, Carsten Jung, and Shayak Sen. “Machine Learning Explainability in Finance: An Application to Default Risk Analysis.” Bank of England Staff Working Paper no. 816, 2019. https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

10Scott M. Lundberg and Su-In Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems 30 (2017): 4765–4774.

Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. New York: ACM, 2016. https://doi.org/10.1145/2939672.2939785.

Credit Fusion and Will Cukierski. “Give Me Some Credit.” Kaggle competition, 2011. https://www.kaggle.com/competitions/GiveMeSomeCredit.

Datta, Anupam, Shayak Sen, and Yair Zick. “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems.” In 2016 IEEE Symposium on Security and Privacy, 598–617. IEEE, 2016. https://doi.org/10.1109/SP.2016.42.

Hofmann, Hans. “Statlog (German Credit Data).” UCI Machine Learning Repository, 1994. https://doi.org/10.24432/C5NC77.

Home Credit Group. “Home Credit Default Risk.” Kaggle competition, 2018. https://www.kaggle.com/competitions/home-credit-default-risk.

Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” In Advances in Neural Information Processing Systems 30 (2017): 4765–4774.

Monetary Authority of Singapore. “Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector.” 2018. https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat.

Yeh, I.-C. “Default of Credit Card Clients.” UCI Machine Learning Repository, 2009. https://doi.org/10.24432/C55S3H.

Table 1. The table shows the discrimination, classification, and error-rate metrics for the main Home Credit XGBoost model on its pre-calibration validation data, alongside threshold outputs and the Brier score.

Metric	Validation result
ROC-AUC	0.7760
PR-AUC	0.2742
Precision at 0.5	0.1803
Recall at 0.5	0.6858
F1 at 0.5	0.2856
Balanced accuracy at 0.5	0.7060
Brier score before calibration	0.1844

Table 2. This table compares independently trained XGBoost models on GMSC, German Credit, and Taiwan data using ROC-AUC, PR-AUC, precision, recall, F1, balanced accuracy, and Brier score on each validation partition.

Dataset	ROC-AUC	PR-AUC	Precision	Recall	F1	Balanced accuracy	Brier
GMSC	0.8701	0.4033	0.2176	0.7855	0.3408	0.7916	0.1405
German Credit	0.7794	0.5955	0.4839	0.7500	0.5882	0.7036	0.1938
Taiwan Credit Card	0.7791	0.5565	0.4750	0.6232	0.5391	0.7138	0.1796

Table 3. This table compares actual default rates with mean predicted probabilities, Brier scores, and expected calibration errors, showing differences between observed and model-estimated risk across the three datasets.

Dataset	Actual default rate	Mean predicted probability	Brier	ECE (10 bins)
GMSC	0.06683	0.32047	0.14051	0.25363
German Credit	0.30000	0.45900	0.19376	0.15900
Taiwan Credit Card	0.22117	0.42334	0.17960	0.20217

Table 4. This table compares validation metrics for GMSC and Taiwan under random and group-aware splits, including discrimination, precision, recall, F1, balanced accuracy, and Brier score.

Dataset	Split	ROC-AUC	PR-AUC	Precision	Recall	F1	Balanced accuracy	Brier
GMSC	Random	0.87012	0.40332	0.21763	0.78554	0.34084	0.79164	0.14051
GMSC	Group-aware	0.86230	0.39217	0.21483	0.77239	0.33616	0.78536	0.14097
Taiwan	Random	0.77910	0.55650	0.47500	0.62320	0.53910	0.71380	0.17960
Taiwan	Group-aware	0.78958	0.57057	0.48287	0.64227	0.55128	0.72240	0.17529

Table 5. This table shows how well the four principles—Fairness, Ethics, Accountability, and Transparency—were applied, highlighting supporting evidence, identifying issues encountered, and indicating the extent of adherence to each principle.

Principle	Evidence and finding	Limitations	Assessment
Fairness	Home Credit gender/age audits and external subgroup audits identify differences in selection rates, TPR, and FPR.	Observational metrics do not establish causal discrimination; some subgroups are small.	Partially addressed

Principle	Evidence and finding	Limitations	Assessment
Ethics	Public dataset provenance, data-quality checks, duplicate review, and leakage-aware preprocessing were considered.	Historical bias, consent, representativeness, and privacy governance are not fully established.	Partially addressed
Accountability	Models, preprocessors, metrics, checkpoints, and audit outputs were saved.	No production governance, formal independent audit, or deployed human-review process exists.	Partially addressed
Transparency	SHAP, performance, calibration, stress, and subgroup metrics are documented.	SHAP is not causal; external probability calibration is weak; uncertainty reporting remains incomplete.	Partially addressed

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)