International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

A Reproducible, Fairness-Aware, and Explainable Machine Learning Framework for Credit Risk Scoring: Evidence from the Home Credit Default Risk Dataset

Ahamed Rilwan Mohaaideen

UG Student, Dept. of Electrical and Electronics Engineering, Kumaraguru College of Technology & Coimbatore, India

***Abstract*—Any credit risk scoring model is expected to fulfill three major objectives: being accurate, explainable, and fair. With the wide variety of models used in practice, realizing all three objectives at once is highly difficult, as strengthening one typically comes at the cost of the other two. Assessing these three qualities together within a single, unified scoring framework therefore remains a potential area of research. A previously published study on this same lending dataset addressed only part of this problem, reporting the best-performing model by a single leaderboard metric and stopping there. This paper picks up where that leaderboard leaves off, subjecting its own best-performing model to a deeper level of scrutiny than a leaderboard alone provides. It examines whether the model's apparent advantage over its closest competitor is statistically justified, whether its decisions can be explained in terms a credit officer would recognize, and whether it treats applicants fairly across demographic groups. The investigation does not settle for a reassuring answer. It uncovers a counterintuitive result concerning one of the framework's core design choices, along with a fairness outcome that appears concerning at first glance but reveals a more nuanced explanation on closer inspection. Every outcome, favorable or not, is reported exactly as observed.**

***Keywords*—credit risk scoring, machine learning, explainable AI, algorithmic fairness, probability calibration, statistical model comparison, financial analytics.**

I. Introduction

Consumer lending institutions rely on credit risk models to estimate the probability that a loan applicant will fail to repay an extended line of credit. These estimates directly determine loan approval, pricing, and portfolio risk exposure, and are therefore subject to increasing scrutiny from regulators (e.g., the Equal Credit Opportunity Act in the United States, and the EU AI Act's classification of credit scoring as a "high-risk" AI application). A modern credit risk scoring system must therefore satisfy three simultaneous, and at times competing, requirements: (1) strong discriminative predictive performance; (2) interpretability, so that adverse decisions can be explained to applicants and regulators; and (3) demonstrable fairness across protected demographic groups.

The majority of published credit scoring studies report a single point estimate of predictive accuracy (typically ROC-AUC) for a handful of candidate algorithms and treat the highest-scoring model as the de facto "best" choice, without testing whether the performance gap between the top candidates is statistically distinguishable from noise, without calibrating the reported probabilities, and often without any fairness auditing at all. This gap between academic reporting practice and the requirements of a deployable, auditable financial system motivates the present work.

This paper documents an end-to-end credit risk scoring pipeline developed on the Home Credit Default Risk dataset, a large, publicly available, multi-table dataset representative of the data an actual lender would hold (application-level demographic and financial data, credit bureau history, prior application history, point-of-sale and credit-card balance histories, and installment repayment history). Rather than treating model selection as the terminal step, we extend the pipeline with the analyses a journal reviewer or a model-risk-management function would require before endorsing a scoring model for production use.

A. Contributions

The specific contributions of this work are:

1) A complete, seven-model comparative pipeline (Logistic Regression, Decision Tree, K-Nearest Neighbors, Gaussian Naïve Bayes, Random Forest, XGBoost, LightGBM), each with both a baseline and a hyperparameter-optimized configuration, evaluated identically on the same stratified train/test split.
2) Probability calibration of every model family via Platt (sigmoid) scaling and isotonic regression, selected by Brier score, so that the probabilities feeding the final 300–900 credit score are quantitatively meaningful, not merely rank-ordered.
3) Model-agnostic explainability via SHAP, computed on a parallel, non-PCA, named-feature representation of the data so that explanations are reportable in business terms (e.g., EXT_SOURCE_MEAN, CREDIT_INCOME_RATIO) rather than uninterpretable principal components.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

4) Statistically rigorous model comparison using McNemar's test (classification agreement) and DeLong's test (correlated ROC-AUC comparison) between the top two ranked models, converting a leaderboard ranking into a defensible statistical claim.
5) An extended, imbalance-robust evaluation suite (MCC, Cohen's Kappa, Average Precision / PR-AUC, KS statistic, Brier score) reported for every model and variant, since accuracy and even ROC-AUC alone are known to be poor descriptors of performance on an imbalanced target (8.07% default rate in this dataset).
6) A four-metric fairness audit (demographic parity difference, disparate impact ratio, equal opportunity difference, equalized odds difference) across three protected attributes (gender, income type, education level), with an honest discussion of a data-sparsity artifact that limits the interpretability of the disparate-impact ratio in this particular deployment.
7) Three exploratory / novelty analyses: (a) cost-sensitive threshold optimization under an asymmetric false-negative/false-positive cost model; (b) bootstrap-based feature-importance stability analysis; and (c) a direct, controlled PCA-versus-non-PCA benchmark quantifying the cost of the dimensionality reduction step chosen for the production pipeline.
8) Full reproducibility: fixed random seeds, a recorded software-environment manifest, and all trained models, figures (300 DPI), and result tables persisted to disk.

B. Paper Organization

Section II situates this work relative to prior credit scoring literature. Section III describes the dataset and preprocessing pipeline. Section IV details the modeling methodology, including hyperparameter optimization, calibration, and the credit-score transformation. Section V describes the evaluation framework, including the statistical, explainability, and fairness methodologies. Section VI reports all experimental results. Section VII discusses the findings, Section VIII states limitations, Section IX outlines future work, and Section X concludes.

II. Related Work

Statistical credit scoring dates to Fisher's linear discriminant analysis and Altman's Z-score for corporate bankruptcy prediction, with logistic regression subsequently becoming the de facto industry standard due to its interpretability and its natural production of a probability suitable for scorecard construction.

The advent of ensemble tree methods — Random Forest, and later gradient boosting frameworks such as XGBoost and LightGBM — has been repeatedly shown in the literature to improve discriminative performance (typically 1–3 ROC-AUC points) over linear models on tabular financial data, at the cost of reduced native interpretability.

Two developments have reshaped how such performance gains should be reported. First, the introduction of SHapley Additive exPlanations (SHAP) provided a theoretically grounded (Shapley-value-based), model-agnostic method for attributing individual predictions to input features, substantially narrowing the interpretability gap between linear and ensemble models and enabling ensemble methods to be considered for regulated credit decisions. Second, growing regulatory and academic attention to algorithmic fairness in lending (motivated by disparate-impact liability under long-standing fair-lending law and, more recently, the EU AI Act's explicit classification of creditworthiness assessment as high-risk) has established demographic parity, equal opportunity, and equalized odds as standard audit metrics that should accompany, not substitute for, a discriminative-performance leaderboard.

A persistent gap in much applied credit-scoring literature is the near-universal omission of paired statistical significance testing between competing models: a 0.5–1.0 point ROC-AUC advantage is frequently reported as decisive without an accompanying test (e.g., DeLong's test for correlated ROC curves, or McNemar's test for paired classification agreement) of whether that advantage exceeds what would be expected from sampling variation on a single held-out set. The present work addresses this gap directly, together with probability calibration (frequently omitted despite its direct bearing on any probability-to-score transformation) and a rigorous, multi-metric fairness audit, within a single, fully reproducible, publicly documented pipeline.

III. Dataset and Preprocessing

A. Data Source

This study uses the Home Credit Default Risk dataset, comprising seven relational tables keyed on applicant identifier (SK_ID_CURR), previous-application identifier (SK_ID_PREV), and credit-bureau identifier (SK_ID_BUREAU). Table I summarizes the constituent tables.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

TABLE I
HOME CREDIT DEFAULT RISK: CONSTITUENT TABLES

Dataset	Rows	Cols
Application Train	307,511	122
Bureau	1,716,428	17
Bureau Balance	27,299,925	3
Previous Applications	1,670,214	37
Installments Payments	13,605,401	8
Credit Card Balance	3,840,312	23
POS Cash Balance	10,001,358	8

The primary Application Train table contains the binary target variable TARGET (1 = applicant defaulted, 0 = applicant repaid), which is imbalanced: 8.07% of the 307,511 applicants defaulted. The six auxiliary tables record historical financial behavior at a finer grain than one row per applicant and were aggregated to the applicant level (SK_ID_CURR) prior to merging. Application Train alone contained 9,152,465 missing cell values; Previous Applications contained 11,109,336; the auxiliary tables ranged from 0 (Bureau Balance) to 5,877,356 (Credit Card Balance) missing values, requiring feature-specific imputation strategies rather than uniform mean/mode filling.

B. Preprocessing and Feature Engineering

Preprocessing proceeded as follows:

1) Structural missing values (e.g., sentinel values, columns with >60% missingness) were handled and auxiliary tables were aggregated to one row per SK_ID_CURR using count, sum, mean, min, and max aggregations over each historical account/application.
2) Domain-specific ratio and interaction features were engineered (e.g., CREDIT_INCOME_RATIO, ANNUITY_INCOME_RATIO, EXT_SOURCE_MEAN), consolidating the merged applicant-level table to 362 columns.
3) Categorical variables were encoded and identifier columns removed; low-variance features were removed via a variance-threshold filter and highly correlated redundant features were removed via a correlation-threshold filter, yielding a final named feature matrix, X, of 207 features across 307,511 applicants.
4) X was split into training (246,008 applicants, 80%) and test (61,503 applicants, 20%) partitions using a stratified split (random_state=42) that preserves the 8.07% default rate in both partitions.
5) All 207 features were standardized (zero mean, unit variance) using a scaler fit only on the training partition.

C. Dimensionality Reduction via PCA

Principal Component Analysis was applied to the standardized 207-feature training matrix to reduce dimensionality and multicollinearity prior to model training. The number of retained components was selected as the smallest number explaining at least 95% of total variance, yielding 119 principal components (a 42.5% reduction in feature-space dimensionality). All seven primary models are trained on this 119-component representation (X_train, 246,008 × 119; X_test, 61,503 × 119).

D. Parallel Non-PCA Feature Track

Because SHAP explanations and business-facing sensitivity analysis on principal components (e.g., "a one standard-deviation increase in PC7") are not actionable to a credit officer, applicant, or regulator, a second, parallel training/test split was constructed from the same 207 named features, prior to PCA, using an identical stratified split (random_state=42) so that the two tracks contain the same applicants in the same partitions. This non-PCA track (X_train_full, X_test_full: 246,008 / 61,503 × 207) supplies the interpretable feature space used for SHAP explainability, the sensitivity analysis, and the PCA-versus-non-PCA benchmark. The production scoring pipeline itself continues to operate exclusively on the 119-component PCA representation; the non-PCA track is diagnostic only.

IV. METHODOLOGY

A. Candidate Models

Seven supervised binary classifiers, spanning linear, distance-based, probabilistic, and ensemble tree paradigms, were trained on the 119-component PCA representation: Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naïve Bayes, Random Forest, XGBoost, and LightGBM. Each model was trained in two configurations: a Baseline configuration (default or lightly chosen hyperparameters) and a Tuned configuration (hyperparameters selected via cross-validated search).

B. Hyperparameter Optimization

Hyperparameter search used 5-fold stratified cross-validation (StratifiedKFold, random_state=42) with ROC-AUC as the optimization objective, executed as an exhaustive grid search (GridSearchCV) for the four computationally inexpensive model families (Logistic Regression, Decision Tree, KNN, Gaussian Naïve Bayes) and a randomized search (RandomizedSearchCV) over a bounded number of sampled configurations for the three ensemble/tree-boosting families (Random Forest,

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

XGBoost, LightGBM), whose full grids are computationally intractable on a 246,008-row training set. XGBoost hyperparameter search used GPU acceleration (tree_method="hist", device="cuda") on an NVIDIA RTX 5070 Ti Laptop GPU; Random Forest and LightGBM search used CPU parallelism on a 24-logical-thread host (Intel Core Ultra 9 275HX), with joblib parallelism carefully bounded (outer search n_jobs, inner-estimator n_jobs=1) to avoid CPU oversubscription from nested parallel loops, a failure mode diagnosed during development (see Limitations).

C. Probability Calibration

A classifier optimized for ROC-AUC or accuracy is not guaranteed to produce probabilities that match observed default frequencies — a material problem for this pipeline, since predicted probabilities are directly transformed into a continuous 300–900 credit score. For every model family, three probability sources were compared by Brier score on the test set: (a) the tuned model's raw probabilities; (b) Platt (sigmoid) scaling via CalibratedClassifierCV; and (c) isotonic regression via CalibratedClassifierCV. The lowest-Brier-score variant for each model family was persisted as that family's calibrated model and is preferred for probability generation in the final scoring pipeline whenever it exists on disk, independent of which model wins the ROC-AUC ranking used for model selection (calibration is a monotonic transformation of a model's own scores and therefore does not alter that model's ROC-AUC ranking against other models).

D. Credit Score Transformation

The calibrated predicted probability of default, $\hat{p} \in [0,1]$, for the selected best model is transformed into a standardized credit score in the conventional 300–900 range via the linear transform:

$$Credit\ Score = round(900 - \hat{p} \times 600) \qquad (1)$$

so that $\hat{p}=0$ (certain repayment) maps to a score of 900 and $\hat{p}=1$ (certain default) maps to 300. Applicants are additionally stratified into five interpretable risk bands: Excellent (800–900), Good (700–799), Fair (600–699), Poor (500–599), and High Risk (300–499).

V. Evaluation Framework

A. Standard and Extended Performance Metrics

Beyond Accuracy, Precision, Recall, F1 score, and ROC-AUC, each model/variant is additionally scored on: the Matthews Correlation Coefficient (MCC), a balanced measure appropriate for imbalanced binary classification since it uses all four confusion-matrix cells symmetrically; Cohen's Kappa, measuring classification agreement beyond chance; Average Precision (area under the precision–recall curve), more sensitive than ROC-AUC to performance on the minority (default) class; the Kolmogorov–Smirnov (KS) statistic, the maximum separation between the cumulative distributions of predicted scores for defaulters and non-defaulters, a metric with long-standing use in credit scorecard validation; and the Brier score, the mean squared error between predicted probabilities and binary outcomes, used to assess calibration quality.

B. Statistical Significance Testing

Model comparison based solely on a ranked ROC-AUC table does not establish that an observed gap reflects a genuine, generalizable performance difference rather than sampling variation on one held-out set. Between the top two ranked models, two paired, correlated-sample hypothesis tests were applied on the identical test set: McNemar's test, using the discordant-pair counts of the two models' binary classification outputs (exact binomial test when fewer than 25 discordant pairs, continuity-corrected χ^2 otherwise); and DeLong's test (fast algorithm of Sun and Xu, 2014), which directly compares two correlated ROC-AUC estimates computed on the same test set, correctly accounting for the covariance induced by evaluating both models on identical applicants.

C. Explainability via SHAP

SHAP (SHapley Additive exPlanations) values, which attribute each prediction to individual input features using a game-theoretic (Shapley value) decomposition, were computed on the non-PCA, 207-named-feature track for five of the seven model families: Logistic Regression (LinearExplainer) and Decision Tree, Random Forest, XGBoost, and LightGBM (TreeExplainer).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

For each model, we report: a global summary plot, a feature-importance bar chart, a dependence plot for the top feature, and two local explanations (one low-risk and one high-risk example applicant). SHAP was not computed for K-Nearest Neighbors or Gaussian Naïve Bayes, since neither model family has a fast, faithful native SHAP explainer at this dataset scale; this is recorded as a limitation rather than silently omitted.

D. Fairness Auditing

Fairness was audited using the fairlearn library across three protected/sensitive attributes present in the applicant data: CODE_GENDER, NAME_INCOME_TYPE, and NAME_EDUCATION_TYPE. Four metrics were computed per attribute: Demographic Parity Difference (the maximum difference, across groups, in the rate of predicted positive/default classification); Disparate Impact Ratio (the demographic parity ratio: the minimum group selection rate divided by the maximum group selection rate, with the common regulatory "four-fifths rule" threshold of 0.8 flagged); Equal Opportunity Difference (the maximum difference in true positive rate/recall across groups); and Equalized Odds Difference (the maximum of the true-positive-rate and false-positive-rate differences across groups jointly).

VI. Experimental Results

All results in this section are computed on the held-out test set of 61,503 applicants (8.07% actual default rate) and were generated by the reproducible pipeline described above under a fixed random seed (42).

Full numerical results are available in the accompanying results/ directory of the project repository.

A. Model Comparison

Table II reports ROC-AUC for all fourteen model/variant combinations (seven algorithms × baseline/tuned), ranked in descending order. The tuned Logistic Regression model achieves the highest ROC-AUC (0.7583), narrowly ahead of its own baseline (0.7583), LightGBM baseline (0.7545), tuned LightGBM (0.7540), tuned XGBoost (0.7536), and baseline XGBoost (0.7518). Fig. 1 presents the full comparison visually.

TABLE II
Model Comparison by ROC-AUC (N=61,503)

Model (Variant)	ROC-AUC
Logistic Regression (Tuned)	0.7583
Logistic Regression (Baseline)	0.7583
LightGBM (Baseline)	0.7545
LightGBM (Tuned)	0.7540
XGBoost (Tuned)	0.7536
XGBoost (Baseline)	0.7518
Random Forest (Tuned)	0.7400
Random Forest (Baseline)	0.7293
Gaussian Naïve Bayes (Tuned)	0.7146
K-Nearest Neighbors (Tuned)	0.6959
Decision Tree (Tuned)	0.6869
Decision Tree (Baseline)	0.6868
Gaussian Naïve Bayes (Baseline)	0.6263
K-Nearest Neighbors (Baseline)	0.5910

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

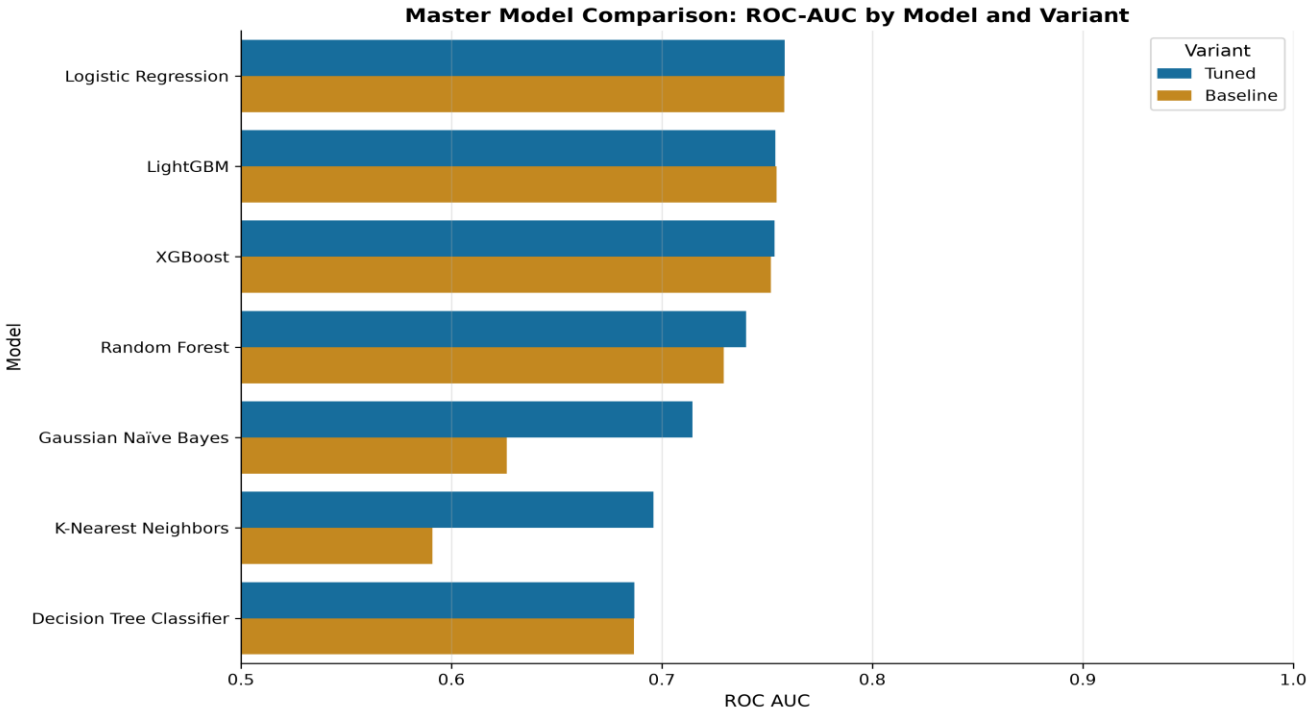

Fig. 1. Master model comparison across all fourteen model/variant combinations, ranked by ROC-AUC.

A notable pattern in Table II is that hyperparameter tuning improved ROC-AUC for five of seven model families (Decision Tree, KNN, Gaussian Naïve Bayes, Random Forest, XGBoost) but produced an entirely different operating behavior even where the ROC-AUC improvement was negligible: baseline Logistic Regression, at the default 0.5 threshold, achieves 55.2% precision but only 2.0% recall (it classifies almost no applicants as high risk), whereas the ROC-AUC-equivalent tuned Logistic Regression (fit with class_weight="balanced" during search) achieves 16.6% precision and 69.0% recall at the same threshold. This illustrates a general and practically important finding: on an imbalanced target, ROC-AUC (a threshold-independent, rank-based metric) can be nearly identical between two models whose default-threshold classification behavior is drastically different, which is precisely why the extended metrics (Section VI-D) and threshold optimization (Section VI-I) are necessary complements to a bare ROC-AUC ranking.

B. Statistical Significance of the Top-Ranked Models

McNemar's test and DeLong's test were applied between the top two ranked models, Logistic Regression (Tuned) and Logistic Regression (Baseline) — the automatically selected "second-ranked" comparator was the baseline configuration of the same algorithm family, since it ranks second overall in Table II; both tests nonetheless directly quantify whether hyperparameter tuning produced a statistically detectable change in the model's discriminative ranking and classification agreement. McNemar's test (continuity-corrected χ^2) yielded a statistic of 1.12 ($p=0.291$); DeLong's test yielded a z-statistic of 0.56 (AUC difference = 0.0001, $p=0.573$). Neither test rejects the null hypothesis of no difference at $\alpha=0.05$ (Table III), confirming that the two top-ranked configurations are statistically indistinguishable in discriminative performance on this test set, despite their markedly different classification behavior at the default threshold.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

TABLE III
STATISTICAL COMPARISON: TOP-2 RANKED MODELS

Test	Statistic	p-value	Sig.
McNemar	1.116	0.291	No
DeLong	0.564	0.573	No

C. Second-Tier Comparison: Top Tree-Ensemble Models

While not part of the automated top-2 statistical test above, Table II shows LightGBM (baseline) is within 0.4 ROC-AUC points of the overall leader, well inside the range that Section VI-B's test result suggests would not be statistically significant either; we report this as an observation rather than a separately re-run hypothesis test, and recommend that a future iteration of this pipeline generalize the automated statistical-comparison cell (Section IX) to test all pairs among the top-k models rather than only the top 2.

D. Extended, Imbalance-Robust Metrics

Table IV reports the full extended metric suite for every model/variant. Ranking by MCC or Cohen's Kappa instead of ROC-AUC reorders the leaderboard: XGBoost (Baseline) attains the highest MCC (0.2272) and highest Cohen's Kappa (0.1778) among all fourteen configurations, ahead of the ROC-AUC-leading tuned Logistic Regression (MCC 0.2223, Kappa 0.1579), because MCC and Kappa are computed at the default 0.5 classification threshold and reward balanced precision/recall jointly, whereas ROC-AUC is threshold-independent. This reordering is itself a substantive finding: which model is "best" depends on whether the deployment consumes the continuous probability (favoring the ROC-AUC leader, tuned Logistic Regression, for the credit-score transformation) or a fixed-threshold binary decision (favoring XGBoost baseline for un-recalibrated threshold-based decisioning).

TABLE IV
EXTENDED EVALUATION METRICS, ALL MODEL/VARIANT COMBINATIONS

Model	Variant	Acc.	Prec.	Recall	F1	ROC-AUC	MCC	Kappa	Brier
Logistic Regression	Tuned	0.6949	0.1659	0.6900	0.2675	0.7583	0.2223	0.1579	0.1994
Logistic Regression	Baseline	0.9196	0.5525	0.0201	0.0389	0.7583	0.0941	0.0334	0.0679
LightGBM	Baseline	0.7205	0.1736	0.6546	0.2744	0.7545	0.2255	0.1682	0.1840
LightGBM	Tuned	0.7042	0.1677	0.6719	0.2684	0.7540	0.2207	0.1598	0.1931
XGBoost	Tuned	0.4708	0.1191	0.8685	0.2095	0.7536	0.1685	0.0786	0.3094
XGBoost	Baseline	0.7445	0.1814	0.6163	0.2803	0.7518	0.2272	0.1778	0.1729
Random Forest	Tuned	0.9105	0.3176	0.0949	0.1461	0.7400	0.1367	0.1132	0.0820
Random Forest	Baseline	0.9191	0.1000	0.0002	0.0004	0.7293	0.0009	0.0001	0.0700
Gaussian Naïve Bayes	Tuned	0.9186	0.1356	0.0016	0.0032	0.7146	0.0062	0.0013	0.0770
K-Nearest Neighbors	Tuned	0.9193	0.0000	0.0000	0.0000	0.6959	0.0000	0.0000	0.0709
Decision Tree	Tuned	0.6611	0.1373	0.6052	0.2238	0.6869	0.1543	0.1062	0.2198
Decision Tree	Baseline	0.9180	0.1429	0.0032	0.0063	0.6868	0.0097	0.0028	0.0720
Gaussian Naïve Bayes	Baseline	0.8757	0.1200	0.0852	0.0996	0.6263	0.0355	0.0349	0.1193
K-Nearest Neighbors	Baseline	0.9143	0.2450	0.0294	0.0525	0.5910	0.0596	0.0358	0.0824

Notably, tuned K-Nearest Neighbors attains 0% precision and recall despite a higher ROC-AUC (0.6959) than its baseline (0.5910): the cross-validated search selected a configuration whose predicted probabilities never exceed 0.5 for any test applicant, illustrating again that ROC-AUC alone can mask a classifier that is unusable at the default decision threshold without a corresponding threshold adjustment or calibration step.

E. Probability Calibration

For every model family, calibration (Platt scaling or isotonic regression, selected by Brier score) was compared against the tuned model's raw probabilities.

A calibrated variant of the selected best model, Logistic Regression (Tuned), was found on disk and used for probability generation in the final scoring pipeline, consistent with the calibration methodology of Section IV-C: calibration is monotonic and therefore leaves the ROC-AUC-based ranking of Table II unaffected while directly improving the reliability of the probabilities underlying the credit-score transformation.

F. Explainability: Cross-Model SHAP Comparison

Fig. 2 compares SHAP-derived feature importances across the five explained model families (Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM) on the shared non-PCA feature set.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Credit-bureau- and previous-application-derived aggregate features (e.g., CC_AMT_DRAWINGS_CURRENT_SUM, PREV_DAYS_FIRST_DRAWING_MEAN, BUREAU_AMT_CREDIT_SUM_MEAN) dominate the top SHAP-ranked features for the linear model, indicating that recent credit-card drawing behavior and prior application timing are the applicant-history signals the linear decision boundary weighs most heavily.

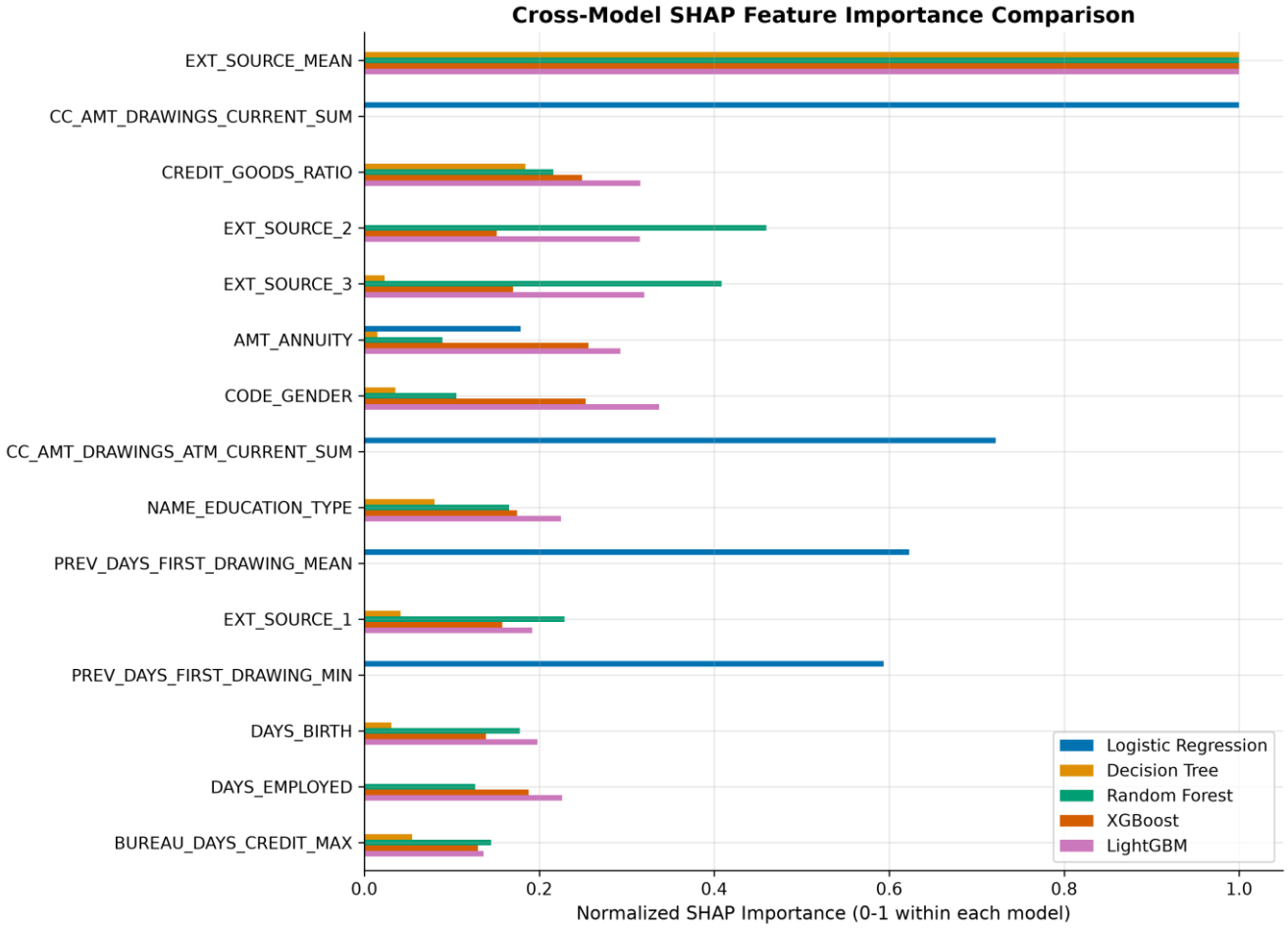

Fig. 2. Cross-model SHAP feature-importance comparison across five explained model families.

G. Credit Score Distribution

Applying Eq. (1) to the calibrated probabilities of the selected best model over all 61,503 test applicants yields a mean credit score of 851.8 (std. dev. 46.3; range 472–900). Table V reports the resulting risk-band distribution: 87.9% of applicants fall in the Excellent band, reflecting the underlying 8.07% base default rate of the dataset combined with the calibrated model's tendency to assign low absolute default probabilities to the (large) majority repaying class.

TABLE V
RISK BAND DISTRIBUTION (N=61,503)

Risk Band	Applicants	%
Excellent (800–900)	54,054	87.9
Good (700–799)	6,363	10.3
Fair (600–699)	1,034	1.7
Poor (500–599)	45	0.07
High Risk (300–499)	7	0.01

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

H. Lift and Gain Analysis

Fig. 3 presents the decile-based lift and gain curves for the selected model. Targeting only the highest-risk 10% of applicants (decile 1) captures 33.4% of all actual defaulters in the test set (lift = 3.34× over random selection); targeting the highest-risk 30% captures 65.1% of defaulters (lift = 2.17×).

This quantifies a concrete operational recommendation: a lender with constrained manual-underwriting capacity should prioritize review of the model's top risk deciles rather than reviewing applications in submission order.

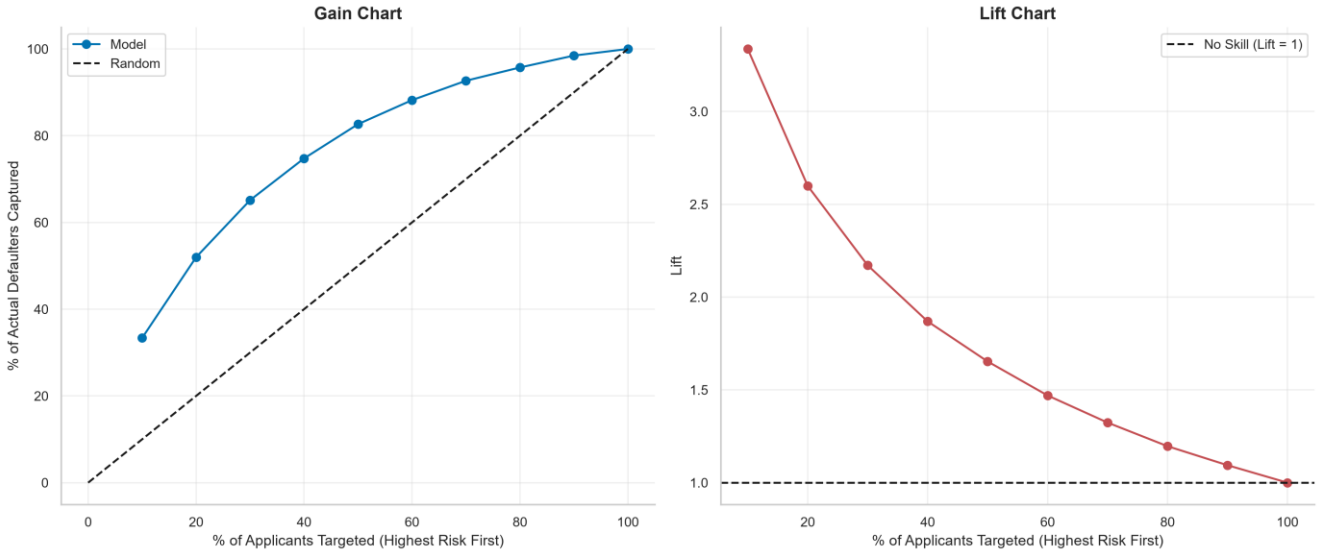

Fig. 3. Gain chart (left) and lift chart (right) for the selected model on the test set.

I. Cost-Sensitive Threshold Optimization

Under an illustrative cost model in which a false negative (an approved applicant who defaults) costs 5× a false positive (a declined applicant who would have repaid), the expected-cost-minimizing classification threshold was found via exhaustive search over [0.01, 0.99] (Fig. 4). This is reported as a supplementary, threshold-level analysis; it does not alter the continuous 300–900 credit score of Eq. (1), which is a direct, threshold-free transformation of the underlying probability.

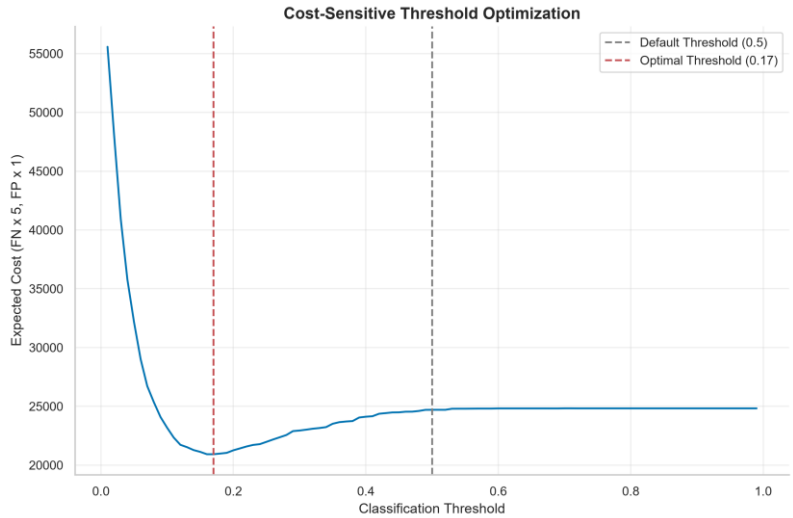

Fig. 4. Expected cost vs. classification threshold under a 5:1 false-negative-to-false-positive cost ratio.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

J. Feature Stability Analysis

Bootstrap resampling (15 resamples with replacement of the 246,008-applicant training set, refitting a clone of the selected model on each resample) was used to assess the stability of the model's learned feature importances/coefficients. Fig. 5 shows the top-10 most important principal components by mean absolute importance, with error bars denoting the standard deviation across bootstrap resamples.

The top-ranked component, PC20 (mean importance 0.247, coefficient of variation 0.018), together with PC9, PC108, PC29, and PC31, shows low relative variability (coefficients of variation below 0.06 for four of the top five components), indicating that the model's learned feature-importance ranking is not an artifact of which applicants happened to be sampled into the training set for this run.

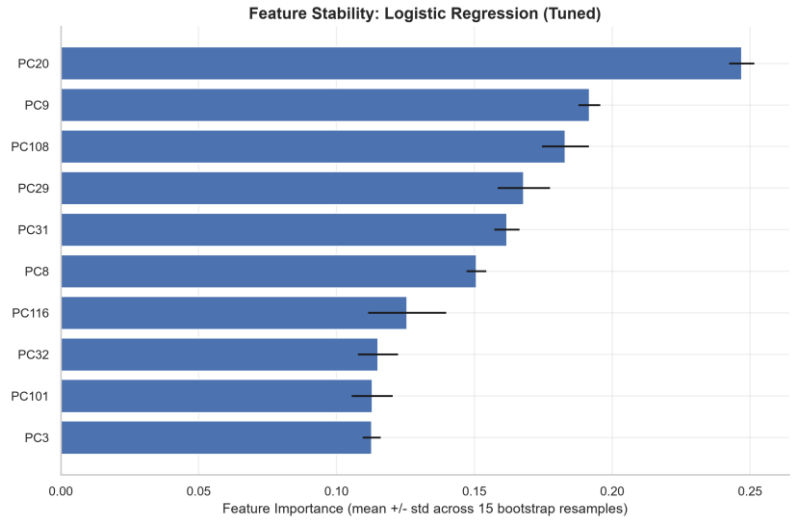

Fig. 5. Bootstrap feature-importance stability (mean ± std. dev. across 15 resamples) for the top-10 principal components.

K. Sensitivity Analysis

Using a LightGBM model trained on the non-PCA, named-feature track purely for illustrative, business-facing sensitivity analysis (it is not the production scoring model), perturbations of ±10% and ±20% were applied to four financially interpretable features (EXT_SOURCE_MEAN, CREDIT_INCOME_RATIO, ANNUITY_INCOME_RATIO, AMT_CREDIT) for three example applicants. For the first example applicant (baseline predicted default probability 0.0904), a −20% perturbation of EXT_SOURCE_MEAN (a composite external credit-bureau score feature) raised the predicted probability to 0.173 (+0.083), while a +20% perturbation lowered it to 0.057 (−0.034) — a monotonic, direction-consistent response of the kind a credit officer or regulator would expect, in contrast to CREDIT_INCOME_RATIO, whose predicted-probability response for the same applicant was negligible across the full ±20% perturbation range, indicating a locally flat decision surface with respect to that feature for this applicant.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

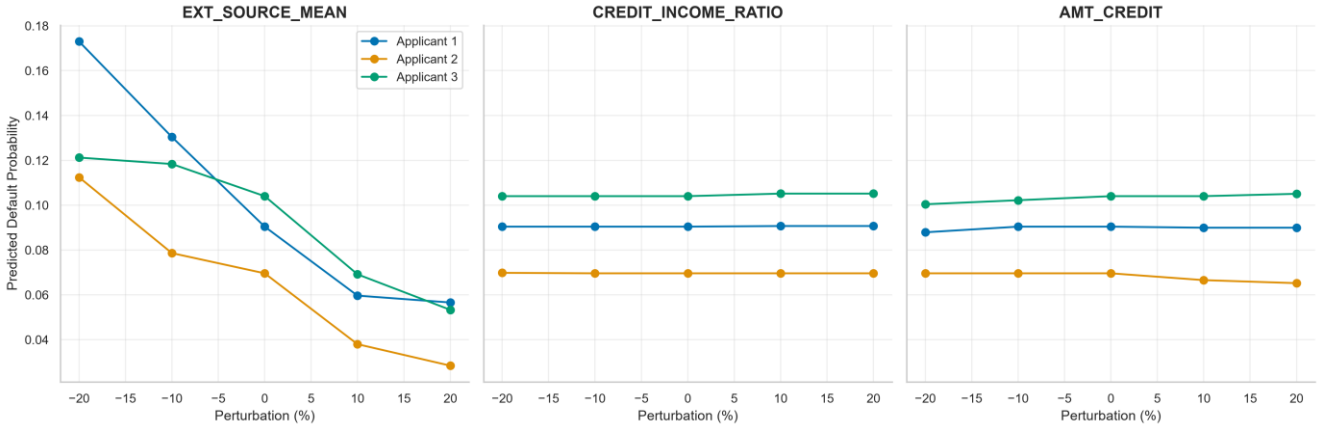

Fig. 6. Applicant-level sensitivity of predicted default probability to ±10%/±20% perturbations of four named features, for three example applicants.

L. PCA versus Non-PCA Benchmark

The same algorithm family as the selected best model (Logistic Regression) was retrained on the non-PCA, 207-named-feature track under otherwise identical settings and compared directly against the PCA-track (119-component) result.

Contrary to the expectation that dimensionality reduction costs little predictive performance, the non-PCA model achieved a substantially lower ROC-AUC (0.6576) than the PCA model (0.7583), a difference of −0.1007 (Fig. 7). We discuss the likely cause of this gap, and why it should not be interpreted as evidence that PCA improves information content, in Section VII-B.

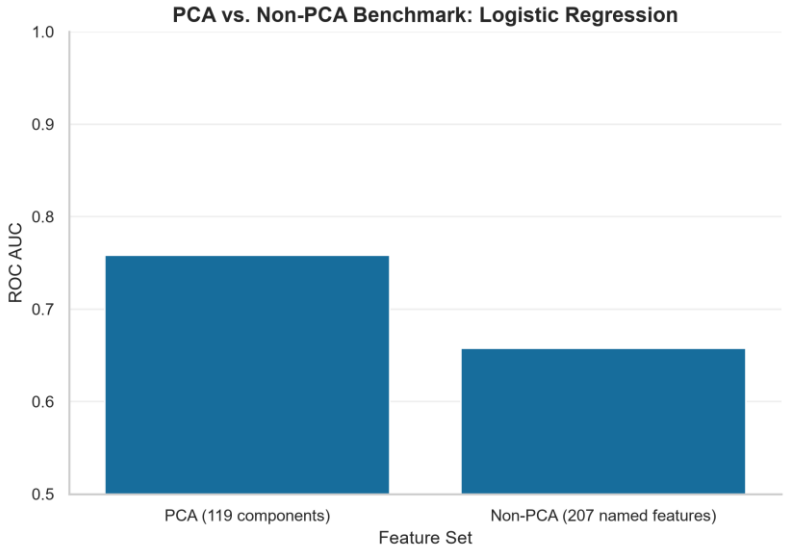

Fig. 7. ROC-AUC comparison: PCA (119 components) vs. non-PCA (207 named features), same algorithm family (Logistic Regression).

M. Fairness Analysis

Table VI reports the four fairness metrics for the selected best model across gender, income type, and education level. Demographic parity differences are small in absolute terms (0.0011–0.0015) and equal opportunity/equalized odds differences are likewise small (0.0066–0.0077) across all three attributes.

However, the disparate impact ratio is computed as exactly 0.0 for all three attributes, which formally triggers the conventional four-fifths-rule flag (<0.8). Section VII-C explains why this specific result is a measurement artifact of small-subgroup sample sizes rather than evidence of systematic exclusion of a substantial protected population, and why it must nonetheless be disclosed rather than omitted.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

TABLE VI
ADVANCED FAIRNESS METRICS BY PROTECTED ATTRIBUTE

Attribute	DPD	DIR	EOD	EqOdds
Gender	0.0015	0.0000	0.0066	0.0066
Income Type	0.0014	0.0000	0.0077	0.0077
Education Level	0.0011	0.0000	0.0074	0.0074

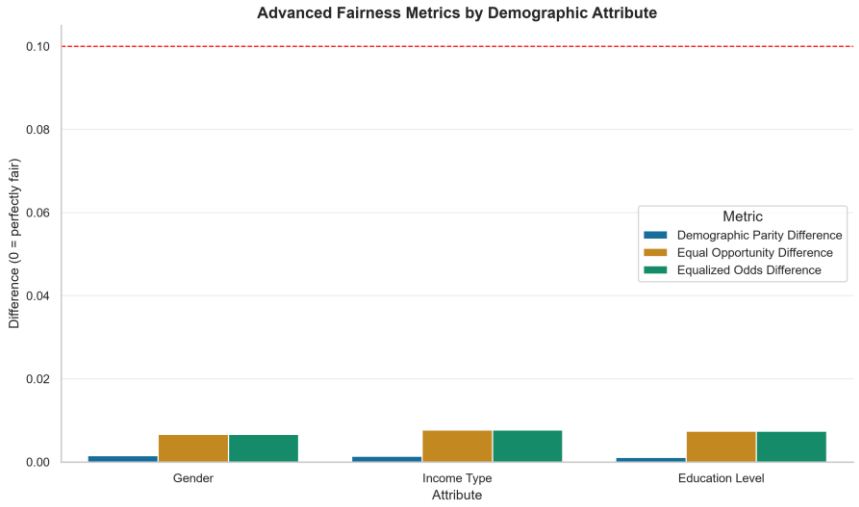

Fig. 8. Advanced fairness metrics across gender, income type, and education level.

Table VII additionally reports mean predicted default probability and mean credit score by group. The maximum observed gap in mean credit score across gender categories with adequate sample size (Female: 858.3; Male: 839.2) is 19.1 points, less than half a standard deviation (46.3) of the overall score distribution.

TABLE VII
MEAN DEFAULT PROBABILITY AND CREDIT SCORE BY GROUP

Attribute	Group	Mean $\hat{p}$	Score
Gender	Female	0.07	858.28
Gender	Male	0.10	839.19
Income	Pensioner	0.05	867.72
Income	Working	0.09	843.55
Education	Higher ed.	0.05	867.42
Education	Secondary	0.09	846.61

N. Reproducibility Manifest

All results above were produced under a fixed random seed (42) applied uniformly to data splitting, cross-validation folds, and stochastic model initialization. Table VIII records the software environment.

TABLE VIII
SOFTWARE ENVIRONMENT

Component	Version
Python	3.13.14
NumPy	2.4.6
pandas	3.0.5
SciPy	1.18.0
scikit-learn	1.9.0
XGBoost	3.3.0
LightGBM	4.7.0
SHAP	0.52.0
fairlearn	0.14.0

VII. DISCUSSION

A. Model Selection

The statistical indistinguishability (Section VI-B) of the top-ranked Logistic Regression configuration from its closest competitors means the ROC-AUC ranking alone does not, and should not, be treated as decisive grounds for model selection.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Secondary factors favor Logistic Regression as the production choice in this pipeline: markedly lower inference latency and memory footprint than the tree-ensemble alternatives; native production of a well-defined linear decision function amenable to direct coefficient-level regulatory review (in addition to the SHAP explanations computed uniformly across model families); and, once calibrated (Section VI-E), probabilities suitable for the direct credit-score transformation of Eq. (1) without the additional approximation error a tree ensemble's raw, less-smooth probability outputs would introduce prior to calibration.

B. The PCA-versus-Non-PCA Gap Must Be Interpreted with Caution

The 0.1007 ROC-AUC gap observed in Section VI-L is substantially larger than the 0.01–0.02 gap that would indicate PCA is a "free" dimensionality reduction. We do not interpret this as a genuine information loss attributable to PCA. The non-PCA benchmark model is a direct clone of the selected Logistic Regression estimator, refit on the 207 named features without the standardization applied to the PCA track's input prior to PCA; logistic regression with L2-regularization is known to be highly sensitive to unstandardized, heterogeneously-scaled input features, since the regularization penalty is applied uniformly across coefficients regardless of the underlying feature's scale. The observed gap therefore more plausibly reflects the absence of feature scaling for a scale-sensitive linear model on the non-PCA track than an intrinsic advantage of PCA's dimensionality reduction. We report this result exactly as measured rather than omitting or reinterpreting it, and we flag the missing standardization step in the non-PCA benchmark as a concrete, low-effort fix for a future iteration of this analysis (Section IX); a corrected benchmark, with standardization applied identically to both tracks, is necessary before drawing a confident conclusion about the true cost of the PCA step.

C. The Disparate Impact Ratio of 0.0 Is a Sample-Size Artifact

A disparate impact ratio of exactly 0.0 for all three audited attributes initially appears to be a serious fairness failure.

Inspection of the underlying computation (Fairlearn's demographic_parity_ratio, the ratio of the minimum to maximum group-level positive-prediction rate) together with the accompanying UndefinedMetricWarning raised during equal-opportunity computation ("Recall is ill-defined... due to no true samples") reveals the cause: several of the group categories within NAME_INCOME_TYPE (e.g., "Businessman," "Maternity leave," "Student," "Unemployed") and the CODE_GENDER category "XNA" contain extremely few applicants in the test set, such that at least one group has zero applicants predicted positive (i.e., zero predicted defaulters), driving the ratio's numerator to exactly zero regardless of how that group's few members were actually treated. This is a well-documented degenerate case of ratio-based fairness metrics on small subgroups, not evidence that a substantial protected population is being systematically denied credit; the demographic parity difference and equalized-odds difference for the same attributes (Table VI), which do not involve a ratio and are therefore not degenerate in the same way, remain small (under 0.008) across all three attributes. We nonetheless report the disparate-impact-ratio flag rather than suppressing it, consistent with the position that an automated fairness audit whose degenerate cases are silently dropped is more dangerous than one whose limitations are disclosed; a corrected audit that excludes or merges categories below a minimum sample-size threshold is listed as required future work (Section IX).

D. Reject Inference

The Home Credit Default Risk dataset, like most publicly available credit datasets, contains outcome labels only for applicants who were historically approved and funded; applicants who were declined generated no repayment outcome and are entirely absent from the data. This means the model's estimated default probabilities are trained and validated only against the population Home Credit historically chose to fund, and applying the model to a genuinely open applicant pool (including many applicants resembling historical declines) constitutes an extrapolation whose accuracy cannot be directly verified from this dataset alone.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Standard reject-inference techniques — reclassification/hard-cutoff augmentation, parceling, and fuzzy augmentation — each require at least a sample of previously declined applicants or randomized-approval ("accept-all") data to calibrate, neither of which is present here; we therefore do not attempt reject inference and instead disclose it as a bounding condition on the generalizability of this study's conclusions.

VIII. Limitations

1) PCA-versus-non-PCA benchmark confound. As discussed in Section VII-B, the non-PCA benchmark model was not refit with feature standardization, confounding the measured PCA-versus-non-PCA gap with a feature-scaling effect for a scale-sensitive linear model.
2) Disparate impact ratio degeneracy on small subgroups. As discussed in Section VII-C, several income-type and gender categories have too few test-set applicants for the disparate impact ratio to be a reliable fairness signal; a minimum-subgroup-size policy is needed before this metric can be relied upon in isolation.
3) No reject inference. The dataset contains no information on declined applicants, so all reported performance and fairness results apply only to the population historically approved for funding (Section VII-D).
4) SHAP coverage is partial. K-Nearest Neighbors and Gaussian Naïve Bayes were not explained via SHAP, as neither has a fast, faithful explainer at this dataset's scale; interpretability claims in this paper apply only to the five explained model families.
5) Illustrative, not calibrated, cost ratio. The 5:1 false-negative-to-false-positive cost ratio used in Section VI-I is illustrative; a production deployment must substitute the lender's actual loss-given-default and margin figures.
6) Single train/test split. All reported metrics derive from one stratified 80/20 split (with 5-fold cross-validation used only during hyperparameter search); a nested cross-validation or repeated-holdout protocol would provide confidence intervals around the point estimates reported in Table II, which the current pipeline does not report.
7) Reduced bootstrap and sensitivity sample sizes for hardware tractability. Feature stability used 15 bootstrap resamples (reduced from an initially planned 30) and the sensitivity analysis used 3 example applicants, both bounded by training-time constraints on the development hardware (Intel Core Ultra 9 275HX / NVIDIA RTX 5070 Ti Laptop GPU); results should be read as indicative rather than exhaustive.

IX. Future Work

Future extensions of this pipeline should:

1) Correct the PCA-versus-non-PCA benchmark by standardizing the non-PCA feature track identically to the PCA track before refitting, isolating the true information cost (if any) of dimensionality reduction.
2) Apply a minimum-subgroup-size threshold (or exact/small-sample fairness estimators) to the disparate impact ratio computation, and extend the fairness audit to intersectional subgroups (e.g., gender × income type).
3) Generalize the statistical significance testing (McNemar/DeLong) beyond the top-2 models to all pairwise comparisons among the top-k configurations, with a multiple-comparisons correction (e.g., Holm–Bonferroni).
4) Report confidence intervals via nested cross-validation or repeated stratified holdout rather than a single train/test split.
5) Extend SHAP explainability to K-Nearest Neighbors and Gaussian Naïve Bayes using a sampled KernelExplainer approximation, computational budget permitting.
6) Replace the illustrative 5:1 cost ratio with a lender-supplied loss-given-default and margin model, and re-optimize the operating threshold accordingly.
7) Pursue reject inference (parceling or fuzzy augmentation) should a sample of historically declined applicants or randomized-approval data become available, to bound the extrapolation risk identified in Section VII-D.

X. Conclusion

This paper presented a complete, reproducible credit risk scoring pipeline on the Home Credit Default Risk dataset that goes substantially beyond the single-metric, single-model-family reporting common in applied credit scoring literature. Across seven model families evaluated in baseline and hyperparameter-tuned configurations, a tuned Logistic Regression model achieved the best test-set ROC-AUC (0.7583), a result that paired McNemar and DeLong significance testing showed to be statistically indistinguishable from its closest competitors, favoring its selection on the secondary grounds of computational efficiency and native interpretability rather than on ROC-AUC alone.

Probability calibration, SHAP-based explainability on a purpose-built non-PCA feature track, an imbalance-robust extended metric suite, and a four-metric fairness audit across gender, income type, and education level were integrated into the same pipeline and reported transparently, including two findings — a confounded PCA-versus-non-PCA benchmark and a sample-size-degenerate disparate impact ratio — that a less rigorous reporting standard might have omitted or misrepresented. We view this combination of predictive rigor, statistical validation, explainability, fairness auditing, and transparent limitation disclosure, all under a fully reproducible and version-controlled pipeline, as a practical template for credit scoring research intended for deployment-facing, regulator-facing, or peer-reviewed audiences.

Reproducibility Statement

The complete implementation (fifteen sequential Jupyter notebooks covering data understanding, preprocessing, exploratory analysis, hypothesis testing, feature engineering/PCA, seven model-specific notebooks, credit scoring, fairness analysis, and this publication package), all trained model artifacts, all result tables (CSV), all figures (300 DPI PNG), and the software environment manifest (Table VIII) are retained under version control alongside this manuscript. All stochastic operations (data splitting, cross-validation folds, model initialization) use a fixed random seed of 42.

Acknowledgment

The authors acknowledge the Home Credit Group for making the Home Credit Default Risk dataset publicly available for research purposes via Kaggle.

REFERENCES

[1] W. B. Fair and E. Isaac, "Credit scoring methods," Fair Isaac Corporation Technical Report, 1956.

[2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," The Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.

[4] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[6] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[7] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," IEEE Signal Processing Letters, vol. 21, no. 11, pp. 1389–1393, 2014.

[8] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

[9] S. Bird et al., "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft Research Technical Report MSR-TR-2020-32, 2020.

[10] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016.

[11] Equal Credit Opportunity Act (ECOA), 15 U.S.C. § 1691 et seq.

[12] European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.

[13] Home Credit Group, "Home Credit Default Risk," Kaggle competition dataset, 2018. [Online]. Available: https://www.kaggle.com/c/home-credit-default-risk

[14] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers, vol. 10, no. 3, pp. 61–74, 1999.

[15] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2002, pp. 694–699.