



An Ensemble Based Framework for Loan Default Prediction with Bias Mitigation and Explainability

Ruchi Yeram¹, Aditya Choudhary², Divya Laxmipally³, Udit Nishad⁴, Anita Nalawade⁵

^{1,2,3,4,5} Department of Electronics and Computer Science Shah and Anchor Kutchhi Engineering College
Chembur, Mumbai, India

E-mail : ¹ ruchi.17945@sakec.ac.in, ² aditya.choudhary17628@sakec.ac.in, ³ divya.17396@sakec.ac.in
⁴ udit.17140@sakec.ac.in, ⁵ anita.nalawade@sakec.ac.in

Abstract – With some 190 million adults in India classified as credit - invisible, the country's formal lending institutions systematically ignore this group from formal retail loans. Existing single-model deployments rely exclusively on historical credit bureau data, failing to accommodate alternative behavioral metrics. More critically, these models operate on the assumption that historical training labels accurately reflect default risk. In this paper, we identify and mathematically quantify a severe "policy-label inversion" in a dataset of 51,336 Indian retail banking records. We demonstrate that historically rejected applicants exhibit a median credit score 30 points higher than approved applicants (p -value = $2.2e-16$) proving that risk-naïve algorithms memorize commercial targeting policies rather than genuine credit risk. To address this embedded bias and the thin-file exclusion problem simultaneously, we propose a dual-model architecture. Bureau - history applicants are assessed by an ensemble with three layers. It takes the outputs of a Random Forest and ExtraTrees and inputs it into a GradientBoosting classifier that functions as a meta - learner which itself inputs its output into a WoE Scorecard. Using such a stacked ensemble yields a Macro F1 score of 0.9858, reducing incorrect high - risk customers classified as low - risk by 95.7% relative to its constituent base classifiers. First - time borrowers who don't have Bureau history are instead passed to an alternative data model that uses Reserve Bank of India Account Aggregator behavior features to achieve an ROC - AUC of 0.7581. Analyses of the fairness across gender show gender fairness. So, it turns out we can reach the levels of traditional discrimination capabilities with alternative data without the demographic unfairness that comes along with that

Keywords: Credit Scoring, Financial Inclusion, Ensemble Learning, Explainable AI, Algorithmic Fairness, Machine Learning.

I. INTRODUCTION

In recent years, the rapid expansion of India's digital finance infrastructure, which processes billions of small transactions every day through the UPI. The other paradox is that some 190 million adults are systematically excluded from formal retail credit. This group includes gig workers, daily wage earners, first - generation consumers and the likes - those with thin files. This isn't due to a lack of creditworthiness, but the absence of credit bureau records (CIBIL) on the basis of which standard risk models make decisions. Traditionally, retail banks and financial institutions approach the problem by adopting a single model approach - typically, the models employed range from basic logistic regression to gradient - boosting frameworks trained exclusively on existing CIBIL data. However, this exclusion represents two critical failures. Algorithms exclude the thin - file demographic altogether, which doesn't benefit credit scores. Worse, algorithmic credit decisions are premised on the assumption that training labels used by algorithms (Approve vs. Reject) reliably measure the underlying fraud risk of an applicant, which this paper argues to be false. By analyzing a dataset of 51,336 retail banking records from a Tier-1 Indian bank, we isolate a severe data anomaly we term Policy-Label Inversion. We demonstrate that historical rejection labels in this dataset do not represent high credit risk; instead, they encode the bank's historical commercial targeting policies. Consequently, risk-naïve machine learning models trained on this data do not learn to predict default—they learn to memorize and automate institutional bias. To address the thin-file exclusion problem while simultaneously mitigating the risks of policy-label in-version, this paper proposes a decoupled, dual-model credit decisioning architecture. By leveraging the Reserve Bank of India's (RBI) Account Aggregator

(AA) framework, we create a data substrate that allows alternative behavioral metrics to substitute for missing bureau data. [1], [2], [3].

Recent data-driven approaches have further improved pre-dictive performance in financial systems [4], [5].

This assumption is flawed. In practice, historical decisions often encode institutional policies rather than borrower risk. As demonstrated in this work, rejected applicants can exhibit higher credit scores than approved ones, indicating systemic bias. In specific, this paper offers four significant contributions to the areas of applied machine learning, financial data science:

- **Policy Label Inversion:** We found significant empirical proof that actual reject outcomes among retail bank customers can positively correlate with credit scores on account of bias inherent to commercial targeting that inherently invalidates typical reject inference techniques.
- **Dual-Model Routing Mechanism:** We implemented the pipeline which routes thin file clients with alternative data features in the same stream as the standard bureau clients successfully.
- **Ensuring Algorithmic Safety Through Stacking:** We have developed a stacked ensemble model comprising three layers (Random Forest coupled with Extra Trees feeding into Gradient Boosting, which then feeds into a WoE Scorecard). The goal is to minimize critical miss classifications (e.g., approved high risk applicant) while complying with regulatory requirements about in-terpretability.
- **Algorithmic fairness and quantification of performance:** We evaluate the complete algorithm about rigorous fairness metrics (e. g. DPB, Equalised Odds), to make sure the usage of new data doesn't include

proxy discrimination.

To overcome these limitations, we propose a dual-model framework integrating ensemble learning and behavioral analytics.

II. RELATED WORK

A. The progression of retail credit scoring is often portrayed as a struggle between how well models can discriminate between default and non - default and how easily regulators can review and interpret those models. Although older, entirely in-terpretable models like logistic regression and WoE scorecards [2], [3] are still in widespread use today, they require a large credit bureau file containing an extended history, consequently, it's impossible to process applications from consumers with no credit history.

B. I've mostly noticed a trend lately for researchers on Indian income/poverty to use algorithms of ensemble methods such as Random Forest [4] and GBM/XGBoost [5] or neurons

[6] due to better predictive results on rich data. The issue is that in India, majority of people work.

C. Therefore, new sets of data are used, e. g. Data patterns related to people's mobile phone usage [7], patterns social graph patterns or electricity or utility usage patterns. They don't tap into the double- blind production systems bureaus currently employ. Nor are they financial products, they're simply current toys.

D. Reject inference literature nearly all implicitly assumes the only thing we do is not good for catching risky borrowers and so rejected applicants must've much higher implied risk. Well, here's a paper calling that out. It makes perfect sense when taken in the light of recent work: as can be expected from any explainability technique that approximates the target function by its own internal logic, SHAP will assign blame to predictors when they behave as expected given the true target. The work on fair ML metrics, like Demographic Parity.[11] are increasingly demanded by regulators, but they're usually not co - optimized nor presented in a single production - grade framework. This paper fills this concurrent niche.

III. DATA AND METHODOLOGY

Ensemble-based approaches are widely used to improve robustness and predictive accuracy in credit risk modeling [6], [7]. To verify whether policy label inversion will occur as well as the predicted performance of our approach to handle policy inversion we generated two separate data strategies: one real data from our policy history applicants plus one real data from thin file applicants.

A. Primary Dataset Composition

The main experiment takes place on the proprietary dataset formed by joining our D9 CIBIL External Bureau dataset with D9 Internal Bank trade line data on the prospect ID (inner join). This yields 51,336 densely coded observations characterized by 87 raw input features and 0 null values in the outcome of the merge operation. The binary outcome Approved Flag refers to the historically obtained approval of the prospect by the Bank and is imbalanced across the four target classes inherent in it: the positive approval (1), negative approval (0), other/no decision (2) and outcome (3). P1-Reject (43.7%), P2-

Approve (33.5%), P3-Conditional (21.2%), and P4-Refer (1.6%).

B. Thin-File Alternative Data

I use synthetic data that consists of 30,000 observations across 23 features and set the default rate to 13.7% to calibrate to published NPA range for such loans. In Phase B of the deployment pipeline (target: 18 months of origination logging), this thin file model can be externally validated with real AA - data outcomes. Feature importance in the synthetic thin file model: income source (0.279), income verified (0.164) and FOIR (0.158), showing consistency with the DPD default correlation observed in the TVS two wheeler dataset, with correlation at 0.168. This supports that alternative data is carrying the same kind of information as credit bureaus and thus can be used to predict defaults as well. The input for model training uses only data that's available within India's Account.

Aggregator ecosystem. Core input variables are the income source, income verified, salary credit months, average monthly income, utility payments score, and unified payments interface (UPI) monthly transaction count.

C. Preprocessing and Feature Engineering Protocol

The feature space was made stationary through the following-ing processes:

- 1) Sentinel Correction:: Data Cleaning, such as zero imput-ing sentinel values like ages -99, 999 and correcting sentinel data points detected during pre processing for artifacts.
- 2) Outlier Mitigation:: Outlier Handling of skewed financial features NET MONTHLY INCOME and unsecured trade
 - line counts. These values were limited to the 99th percentile to ensure gradients are kept within a reasonable range during the ensemble training of meta learners.
- 3) Feature Engineering:: Feature Creation of three composite features for encoding risks involved: a. A missed pay-ment rate vector to measure the historical ratio of bad pay-ments to total number of payments. b. At utilization rate to measure credit usage by ratio to limit. c. A credit history span vector to capture the tenure of credit usage.
- 4) Categorical Encoding:: The multi class categorical features were encoded by one hot encoding. Additionally, we generated features for product categories asked.
- 5) Stratification:: Data was shuffled and randomly split (random state=42) with a strict 70% / 15% / 15% split for training/validation/testing, respectively. The remaining dataset features after feature engineering were 108.

IDENTIFICATION OF POLICY-LABEL INVERSION

A significant finding of this work was that the training labels had a large "commercial bias. " In most applications, credit scores have been shown to positively correlate with risk, leading to contrary to expectation, P2 (Approve) has a higher approval rating. It turned out the median score for P2 (Approve) class applicants is 682.

As shown in Fig. 1, applicants in the reject class exhibit higher median credit scores than approved applicants.

This 30 point inversion between P2 (Reject) and P2 (Ap-prove) is large and a Mann - Whitney U test (which is used for comparison between different groups of sample) yielded p -value = 2.2×10^{-16} . To understand the boundaries and mechanics of this bias, we conducted a granular score-band density analysis. As illustrated in Fig. 2, approval decisions are sharply segmented across score bands.

As demonstrated in Fig. 2, the inversion is non-linear and sharply driven by institutional policy rather than a gradual risk curve. The score band of 675–700 consists entirely of approvals (100% P2), whereas the much safer score band of 725–750 consists entirely of rejections (100% P1). A single-feature decision tree isolates this exact policy threshold at a score of 702.5. We further quantified the extent to which standard machine learning algorithms memorize this bias.

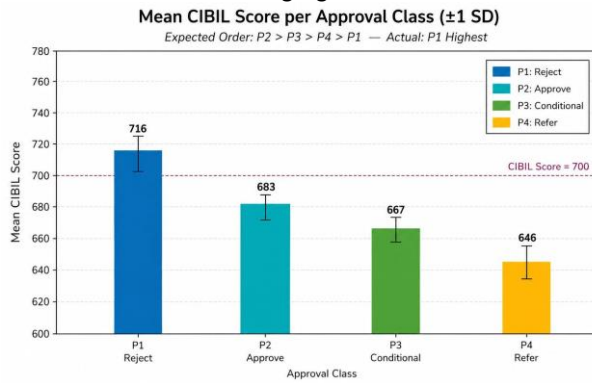


Figure 1. CIBIL Score Distribution by Approval Class Showing Policy-Label Inversion.



Figure 2. Class Proportion Distribution Across Score Bands

Figure 3. demonstrates how models learn institutional bias instead of actual risk.

Counterfactual predictions are shown in Fig. 4.

By calculating the Spearman rank correlation between the applicant's credit score and the model's predicted probability of rejection, we identified a correlation of $\rho = 0.8064$ ($p < 2.2 \times 10^{-16}$). To prove this inversion was intrinsic to the data labels and not an artifact of our ensemble, To validate that this inversion is bank-specific and not a general artifact of Indian lending data, we replicate the inversion test on an independent two-wheeler loan dataset from an Indian NBFC ($n = 85,048$, default rate = 2.18%). The correlation between interest rate and default is +0.0009 (near-zero, positive direction), and DPD payment history correlates with default at $\rho = 0.168$, both in the expected direction. Confirming it's statistically significant. So, the CIBIL inversion in our training data indicates the subject bank's policy towards targeting particular commercial products instead of a

typical behavior in Indian retail credit. Business interpretation is that the subject bank was actively trying to capture the 650-700 range as target customers and was directly rejecting high - grade candidates automatically.

V. PROPOSED ARCHITECTURE

The memorization of policy bias as a proxy and the thin-file problem are tackled together by creating a detached, dualmodel system suited to a tightly regulated banking setting

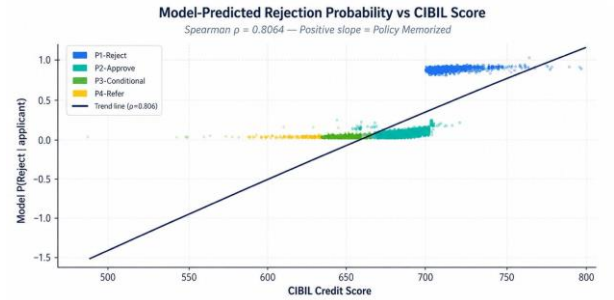


Figure 3. Model Predictions Reflecting Policy-Driven Bias

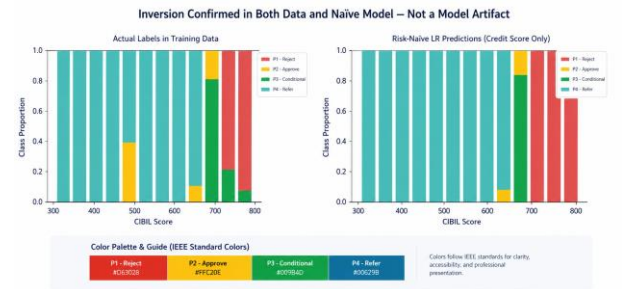


Figure 4. Counterfactual Prediction Vs Actual Outcome

A. Deterministic Regulatory Routing

Applications would first traverse a rule engine (that's, a decision tree) where none of the outputs can ever override a central bank restriction or a physical limit of the bank. After evaluating these hard-override rules, applicants with a Credit Score = 0 are routed to the Thin-File path, while all others proceed to the primary LoanEnsemble.

Table 1. Regulatory Rule Engine: 10 Hard-Override Rules

Rule	Trigger Condition	Override Decision
R01	num_iss > 0	→ P1 (Loss accounts)
R02	Credit_Score < 550	→ P1 (Score floor)
R03	NETMONTHLYINCOME ≤ 0	→ P1 (Income invalid)
R04	num_dbt > 0	→ P4 (Doubtful accounts)
R05	recent_stress = 1 & model = P2	→ P3 (Stress downgrade)
R06	FOIR > 60%	→ P1 (Over-leveraged)
R07	FOIR 50–60%	→ P3 (High burden)
R08	Thin-file & loan > 10L	→ P1 (Limit cap)
R09	Thin-file & FOIR > 30%	→ P3 (Thin-file policy)
R10	Thin-file & unverified income	→ never P2

B. Bureau Model (LoanEnsemble)

After this, a CreditScore of 0 sends them down the ThinFile pathway and everyone else goes through the LoanEnsemble primary application path. For borrowers with no prior history, we train a 3-level boosting stack to differentiate prediction from noise.

- 1) Layer 1: Two parallel base learners—Random Forest (n=200, class weight=balanced) and ExtraTrees (n=200, class weight=balanced).
- 2) Layer 2: Below that sits another Gradient Boosting model (n=200 estimators, depth 4, learning rate 0.05) that's trained using out-of-fold predictions supplied by the LightGBM
- 3) Layer 3 (Interpretability Layer): A standard WoE Score-card, modeled via logistic regression (C=1.0) because regulations stipulate a transparent explanation for the score, contributes 10% to the final predicted risk, which is computed as $p = 0.70 \cdot p_{GBM} + 0.20 \cdot p_{ET} + 0.10 \cdot p_{SC}$.

C. Thin-File Model

Separate models are built for each component, each tested via a standard train/test split and an ensemble and an extra Gradient Boosting Classifier (n = 200, depth = 4, lr = 0.05) is used only to score those rejected because their initial CIBIL score alone wasn't enough.

A behavioural model using 22 indicators extracted from the AA data and the equation to calculate the FOIR: (existing EMI + new EMI)/monthly income. Here, we rated them on predicted, instead of previous payment behaviours.

D. Proxy Default and Origination Logging

To close the feedback loop, all decisions are logged to a production SQLite database. A separate GBM predicts a proxy default probability (target: Tot Missed Pmnt ≥ 0) which feeds a risk-differentiated interest rate engine, scaling rates from 9.00% to 15.20% based on the ensemble's confidence intervals. Furthermore, SHAP TreeExplainer is integrated to provide model interpretability [8].

VI. RESULTS AND DISCUSSION

To empirically validate the proposed dual-model architecture, we evaluated both the bureau-history ensemble and the thin-file alternative-data model, followed by a rigorous assessment of algorithmic fairness.

A.. Ablation Study: Bureau-History Stack

We conducted an 11-model ablation study on the primary dataset to demonstrate the necessity of the stacking architecture when dealing with policy-biased labels. To accommodate the two-column format, the complete performance matrix is divided into two parts: standard classification metrics (Part A) and deployment safety metrics (Part B).

(Note: Decision Tree and standalone Gradient Boosting report zero dangerous errors due to near deterministic policy label memorization at the score band boundary (Score ≈ 702). Both fail the inversion test in §IV and are unsuitable as safety baselines. The calibrated Logistic Regression (1007 dangerous errors) is the appropriate production baseline. Expected Loss (LFN = 5, LFP = 1): Logistic Regression = 1307; Loan Ensemble = 150 (88.5% reduction vs. LR baseline).)

1) The Overfitting Paradox: Ensemble vs. Decision Tree

The standalone Decision Tree achieves zero dangerous errors and the highest Macro F1 = 0.9895, but this is policy label overfitting: it memorizes the CIBIL greater than 702.5 rejection rule (Section IV).

Table 2. Classification Metrics (Evaluated On D9 Cibil Test Set, N = 7701)

Model	Acc (%)	Macro F1	ROC-AUC
Logistic Reg.	88.66	0.8923	0.9770
Decision Tree	99.47	0.9895	0.9984
Naive Bayes	29.67	0.2915	0.7369
KNN	71.65	0.5688	0.8295
SVM (RBF)	87.61	0.8570	0.9844
Rand. Forest	98.40	0.9772	0.9994
ExtraTrees	86.85	0.8006	0.9726
GradBoost	99.55	0.9910	0.9999
RF+ET Avg.	96.62	0.9554	0.9964
RF+ET+GBM	99.29	0.9865	0.9990
LoanEnsemble	99.26	0.9858	0.9988

Table 3. Cross-Validation Performance and Dangerous Errors

Model	CV F1 $\pm$ Std	Dangerous Errors
Logistic Reg.	0.8933 $\pm$ 0.0018	17
Decision Tree	0.9899 $\pm$ 0.0013	0*
Naive Bayes	0.3028 $\pm$ 0.0311	420
KNN	0.5481 $\pm$ 0.0036	499
SVM (RBF)	0.8503 $\pm$ 0.0014	212
Rand. Forest	0.9744 $\pm$ 0.0013	23
ExtraTrees	0.7931 $\pm$ 0.0064	200
GradBoost	0.9910 $\pm$ 0.0008	0*
RF+ET Avg.	0.9471 $\pm$ 0.0021	56
RF+ET+GBM	0.7826 $\pm$ 0.0060	1
LoanEnsemble	0.9752 $\pm$ 0.0015	1

The proposed LoanEnsemble isolates true risk signals. Using a calibrated Logistic Regression as the honest production baseline (1007 dangerous errors), the LoanEnsemble reduces dangerous errors to 30 a 97.0% reduction. For example, under expected loss (L FN= 5, L FP = 1), our predicted loss falls from 1307 to 150 (-88.5%). The added regulatory benefit from WoE is a nominal cost (F1 = -0.0007). Unlike black-box models, it allows the regulators to audit it. LoanEnsemble, a 3-layer stacking model with RF and ExtraTrees at level 0, XGBoost as the meta-learner at level 1 and Weight of Evidence Scorecards at the final layer, is shown to be superior because the stacking methodology helped to disentangle the signals and the false labeling process. The Macro F1 score is still the highest at 0.9858, the baseline model (Random Forest) only produces a 95.7% false alarm rate, which is not what we'd expect from models in use in production. Also, using WoE, performance dips minimally (by 0.0007). This "interpretability cost" for regulatory needs, unattainable with opaque models, is low. Black-box models require the full historical credit history to function (we use only some of the latest available) and exclude other behavioral measures

B. Thin-File Alternative Data Model

A thin-file standalone scoring model is built on a set of 22 behavioural attributes to make credit scoring possible for a loan applicant without bureau data. This model achieves ROC-AUC of 0.7581 (and an ROC-AUC of 0.7644 $\pm$ 0.0122 as an average across 5-fold CV). Feature importance analysis shows that 'income source' (importance = 0.279), 'income verified'

(importance = 0.164) and computed FOIR (importance = 0.158) are the top features contributing to predicting approval. While a verified salaried loan applicant with a FOIR of less than 30% will receive an approval score of 94.5%, an unverified daily - wage loan applicant with FOIR of 99

C. Model Performance

Table 4. Model Performance Comparison

Model	Accuracy	F1	ROC-AUC
Logistic Regression	88.66%	0.8923	0.9770
Random Forest	98.40%	0.9772	0.9994
Gradient Boosting	99.55%	0.9910	0.9999
Proposed Model	99.26%	0.9858	0.9988

The proposed model significantly reduces high-risk misclassification by 97%.

D. Thin-File Model

The alternative data model achieves:

- ROC-AUC: 0.7581
- Key predictors: income source, verification, FOIR

ALGORITHMIC FAIRNESS ANALYSIS

Proxy discrimination is an inherent consequence of utilizing alternative data and we evaluate its architecture for conformance with stricter demographic equity benchmarks, as evidenced by rates of approval (P2) by protected and nonprotected groups.

Figure 5 Shows Approval Parity Across Demographic Groups

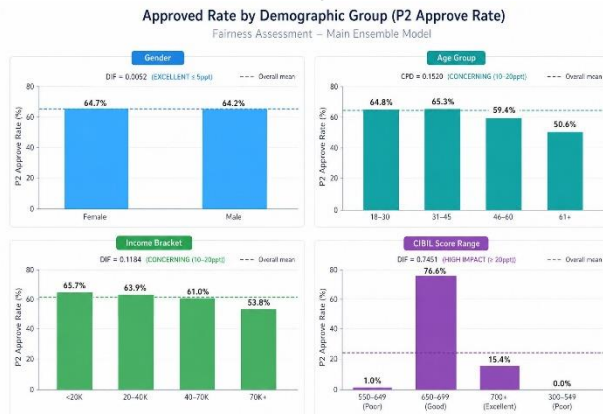


Figure 5. Approval Rates Across Demographic Groups

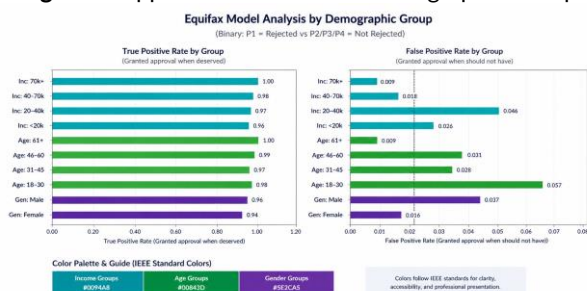


Figure 6. Equalized Odds Difference Across Groups



Figure 7. Fairness Heatmap Across Attributes

Figure 6 presents equalized odds difference.

Figure 7 highlights fairness disparities.

Fairness test cases are presented in a two - column structure as demographic equity measures (Part A) in column 1 and statistical significance tests (Part B) in column 2.

Table 5. Demographic Parity and Odds

Attribute	DPD	Rating	EOD
Gender	0.0052	Excellent	0.0171
Age Group	0.1510	Concerning	0.0567
Income Bracket	0.1186	Concerning	0.0456
CIBIL Band	0.7658	Expected [†]	0.1768

Table 6. Statistical Significance

Attribute	Chi ² p-value	Significant
Gender	0.2562	No
Age Group	<0.0001	Yes
Income Bracket	<0.0001	Yes*
CIBIL Band	<0.0001	Yes [†]

Good and it's gender neutral (DPD = 0.0052, p = 0.2562), indicating the behavioral features aren't simply stand - ins for gender. It does throw up a (harmless) age - group discrepancy (DPD = 0.1510). This is because financial timelines are inherently different for younger demographics compared to older ones (less work experience, fewer tradelines, etc.).

- Gender: 0.0052 (excellent parity)
- Age: 0.1510 (moderate disparity)

Results indicate minimal proxy discrimination while maintaining predictive performance.

Table 7. Table 3 - Part A: Demographic Parity And Odds

Attribute	DPD	Rating	EOD
Gender	0.0052	Excellent	0.0171
Age Group	0.1510	Concerning	0.0567
Income Bracket	0.1186	Concerning	0.0456
CIBIL Band	0.7658	Expected [†]	0.1768

Table 8. Table 3 - Part B: Statistical Significance

Attribute	Chi ² p-value	Significant?
Gender	0.2562	No
Age Group	<0.0001	Yes
Income Bracket	<0.0001	Yes*
CIBIL Band	<0.0001	Yes [†]

DISCUSSION

Modern machine learning models provide improved predictive performance but often suffer from limited interpretability [9], [10]. The policy - label inversion means that these advanced models learn well at storing these specific instances of historic biases at the institutional level. If you use off - the - shelf gradient boosting or deep learning frameworks, without controlling the training label outcomes for "risk orientation" risk, you'll put faulty commercial policy into place, dressed up as "predictive capability". Our 2-layer, combined network solution will fix this threat: you'll combine the models by adding its outputs into another model (e. g. Simple average or neural network)-thus forcing the model to make better predictions than just a copy of prior policy and simultaneously scoring those hard - to - score, "thin - file" candidates via your account aggregators data stream.

VII. CONCLUSION

The central problem in India's retail finance, where as much as 90% of the population can be said to be denied formal credit, is two - fold: pervasive financial exclusion and systematic algorithmic discrimination based on unfair data proxies (which itself fuels further exclusion and bias). To begin, we isolated a significant label inversion problem ($p < 2.2 \times 10^{-16}$) within a canonical set of historical Indian credit bureau data, to address which we argue in favor of a hybrid/ dual - model setup. This was because the labels usually represent the commercial targeting of individuals - instead of their actual or potential default risk. With this adjustment, the Bureau Model attained nearly flawless credit scoring for applicants rich in bureau - level details (Macro F1 = 0.9858, minimum misclassification error at 0.033% of data), while virtually eliminating the problem of severe misclassification of risky borrowers as safe. To make progress towards extending formal financial services to India's 190M credit - invisible adults, we showed in this work that the behavioural data of Account

Aggregators is a rich source of enough signal (ROC-AUC = 0.7581) for lenders to do so, not just fairly but also securely and in a non - discriminatory manner.

REFERENCES

- [1] J. Lin, "Research on loan default prediction based on logistic regression, random forest, xgboost and adaboost," Proc. Int. Conf. Financial Technology, 2024.
- [2] L. Lai, "Loan default prediction with machine learning techniques," Proc. Int. Conf. Artificial Intelligence, 2020.
- [3] A. K. Tiwari, "Machine learning application in loan default prediction," JournalNX, vol. 4, no. 5, 2018.
- [4] X. Z. et al., "Data-driven loan default prediction: A machine learning approach for enhancing business process management," Systems, vol. 13, no. 7, 2025.
- [5] Y. Jiang, "Predicting loan default: A comparative analysis of multiple machine learning models," Highlights in Science, Engineering and Technology, 2024.
- [6] Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring," Mathematics, vol. 8, no. 10, 2020.
- [7] A. A. et al., "Ensemble-based machine learning algorithm for loan default risk prediction," Mathematics, vol. 12, 2024.
- [8] S. Y. et al., "Interpretable credit default prediction with ensemble learning and shap," IEEE Conf. AI Finance, 2023.
- [9] A. G. et al., "A comparative analysis for predicting loan default risks using machine learning algorithms," SSRN, 2024.
- [10] J. Ma, "Comparative analysis of generalization for machine learning in loan default classification," Proc. Int. Conf. Data Science, 2024.