

PsyFint: Leveraging Psychological Profiling for Enhanced Credit Risk Assessment

Erdenebileg Batbaatar

Optimal NMax LLC

Ulaanbaatar, Mongolia

erdenebileg.b@optimal.mn

Byambajargal Shijir

Optimal NMax LLC

Ulaanbaatar, Mongolia

byambajargal@optimal.mn

Nandin-Erdene Bayarmagnai

Optimal NMax LLC

Ulaanbaatar, Mongolia

nandinerdene@optimal.mn

Gantulga Jargalsaikhan

Optimal NMax LLC

Ulaanbaatar, Mongolia

gantulga.j@optimal.mn

Khuslen Galbaatar

Optimal NMax LLC

Ulaanbaatar, Mongolia

khuslen@optimal.mn

Telmen Enkhtuvshin

Optimal NMax LLC

Ulaanbaatar, Mongolia

telmen.e@optimal.mn

Khosgarig Altantsatsralt

Optimal NMax LLC

Ulaanbaatar, Mongolia

khosgarig@optimal.mn

Mandukhai Ganbat

Optimal NMax LLC

Ulaanbaatar, Mongolia

mandukhai@optimal.mn

Abstract—Conventional credit scores focus on financial history and a few demographics while overlooking behavioral heterogeneity. We introduce PsyFint, a credit risk framework that augments standard underwriting with a *Psychological Composite Score* (PCS) built from seven validated subscales. Two auxiliary instruments (mandatory items and attention/lie checks) are used only to ensure response quality. On 59,590 borrowers, integrating PCS into a calibrated LightGBM improves ROC-AUC from 0.68 to 0.71 and reduces the Brier score from 0.167 to 0.161. At a cost-sensitive operating point, specificity stays near 0.90 while recall for high-risk borrowers increases relative to a no-PCS baseline. SHAP analysis ranks PCS subscales alongside repayment-history variables. We report ROC, PR, and KS curves, calibration via expected calibration error, and subgroup diagnostics by gender and education, and we document governance practices for consent, data minimization, and monitoring. The results show that carefully measured psychological traits provide orthogonal and interpretable signal that improves discrimination and calibration, which can support more inclusive and transparent lending.

Index Terms—credit scoring, psychological factors, behavioral finance, machine learning, risk assessment, responsible lending

I. INTRODUCTION

Retail credit scorecards rely primarily on bureau histories and a compact set of demographics that proxy a borrower’s *capacity to repay* [1]. They often underweight *propensity to repay*, which captures behavior under financial stress, impulse spending, and competing obligations. This imbalance contributes to avoidable misclassification, particularly for thin-file or new-to-credit applicants [2]–[4]. To address this gap, we target propensity to repay directly using brief, validated psychological measures that are practical for underwriting workflows.

Nonlinear machine learning improves rank ordering and can capture interaction effects that linear scorecards miss [5]. Deployed systems, however, need calibrated *probabilities of default* (PDs) rather than raw margins, which motivates post-hoc calibration and reliability assessment [6], [7]. In parallel, the rise of alternative data has expanded coverage while raising explainability, privacy, and governance concerns;

production systems must manage interpretability, technical debt, and dataset shift [4], [8]–[10].

Behavioral science offers an interpretable complement to traditional underwriting. Validated constructs such as financial responsibility, rule obedience, impulse control, saving attitude, financial literacy, and locus of control are associated with delinquency and default when measured carefully [2], [3]. Practical use requires response-quality safeguards, including attention or lie checks and consistency screens [11].

We present PsyFint, a behavior-aware scoring framework that augments underwriting data with a compact *Psychological Composite Score* (PCS) built from seven brief subscales and supported by response-quality instruments. The pipeline is designed for deployment and includes quality control, leakage-aware temporal splits, probability calibration, cost-sensitive thresholding, explainability via *SHapley Additive exPlanations* (SHAP), subgroup diagnostics, and monitoring for drift and reliability [6]–[10], [12].

Research questions. We ask: (RQ1) does PCS add discrimination and calibration lift over strong baselines; (RQ2) is the signal complementary to repayment-history variables across thresholds; (RQ3) do calibrated probabilities with PCS improve cost or error trade-offs at realistic operating points; and (RQ4) do explanations and subgroup gaps meet deployment standards.

Contributions. This work provides: (i) a brief validated psychometric instrument with explicit quality control and a standardized PCS; (ii) large-scale evidence that PCS supplies orthogonal and interpretable lift in ranking and calibration; and (iii) a deployment blueprint with calibrated probabilities, cost-aware policy, SHAP reason codes, fairness analysis, and production monitoring.

II. RELATED WORK

A. Traditional Credit Scoring

Classical retail scorecards are typically built with logistic regression and Weight-of-Evidence binning to ensure transparency, monotonicity, and stability across cycles [1]. This

additive structure supports governance and reason codes but can miss interactions, nonlinearities, and behavioral heterogeneity, which is especially limiting for thin-file or new-to-credit borrowers.

B. Machine Learning for Credit Risk

Tree-based ensembles such as Gradient Boosting and LightGBM improve rank ordering by capturing higher-order structure and heterogeneous effects [5]. For deployment, probability outputs must be well calibrated rather than relying on raw margins, which motivates post-hoc calibration and reliability assessment [6], [7]. Under class imbalance, *precision–recall (PR)* analysis provides a more informative view of operating trade-offs than ROC alone [13].

C. Psychological and Behavioral Predictors

A growing literature links psychological constructs to financial behavior and default risk. Constructs include self-control around impulse spending, financial responsibility, saving attitude, financial literacy, rule obedience, and locus of control [2], [3]. Practical use requires safeguards for response quality, such as attention or lie checks and consistency screens, to reduce satisficing and gaming [11].

D. Alternative Data, Explainability, and Governance

Digital footprints and other alternative data can expand coverage and add predictive signal, while raising questions about privacy, portability, and auditability [4]. Model explanations are needed for review and disclosure, and SHAP-based methods provide local and global attributions suited to tree models [8], [14]. Reliable production systems also require controls for technical debt and dataset shift [9], [10] and fairness analysis with calibration or threshold adjustments when subgroup gaps are material [12], [15].

III. DESIGN OF *PsyFINT*

A. System Overview

PsyFint augments standard underwriting with validated psychological profiling in a deployment-oriented pipeline:

- 1) **Data collection:** in-form attention and lie checks with minimal metadata for audit [11].
- 2) **Quality control:** hard filters and optional down-weighting for low-quality responses.
- 3) **Feature construction:** psychological subscales and PCS, operational or financial variables, and digital or demographic signals, all timestamped at decision time t .
- 4) **Modeling:** LightGBM as the primary learner with classical baselines and class-imbalance handling [5].
- 5) **Calibration and policy:** Platt or isotonic calibration with cost-sensitive thresholding [6], [7].
- 6) **Explainability, fairness, and governance:** SHAP-based reason codes, subgroup diagnostics, drift monitoring, and access control [8]–[10], [12], [14], [15].

Algorithm 1 PCS pipeline

- 1: Reverse code negatives and compute subscale means x_{ik} .
- 2: **if** completion $< \rho$ **then**
- 3: $x_{ik} \leftarrow \text{NaN}$
- 4: **end if**
- 5: Standardize on training folds to obtain z_{ik} .
- 6: $\text{PCS}_i \leftarrow \sum_{k=1}^7 z_{ik}$.
- 7: Apply QC: exclude on LD or MIC fail; optional sample weight w_i .

B. Psychometric Instrument and Scoring

Constructs and Items: Seven brief Likert subscales are used: *Money Management (MM)*, *Emotional or Impulse Purchasing (EP)*, *Financial Responsibility (FR)*, *Saving Attitude (SA)*, *Financial Literacy (FL)*, *Rule Obedience (RO)*, and *Locus of Control (LoC)*. Items use a 5-point scale from 1 (Strongly Disagree) to 5 (Strongly Agree). Negatively keyed items are reverse coded.

Validity Instruments: Two auxiliary blocks are administered but excluded from PCS aggregation: (i) **Mandatory Items Check (MIC)** to detect acquiescence or contradictions, and (ii) **Lie or Attention Detection (LD)** with blatant true or false items and timing heuristics [11].

Handling Missing or Partial Responses: For subscale k , if completion is below a threshold ρ (for example, 70%), the subscale is set to missing. Otherwise the subscale score is the mean of answered items. We avoid item-level imputation and aggregate at the subscale level.

C. PCS Construction and Normalization

Let x_{ik} denote the average raw score for borrower i on subscale k after reverse coding. Standardize on the *training* set:

$$z_{ik} = \frac{x_{ik} - \mu_k^{(\text{train})}}{\sigma_k^{(\text{train})}}. \quad (1)$$

Define the **Psychological Composite Score (PCS)**:

$$\text{PCS}_i = \sum_{k=1}^7 z_{ik}. \quad (2)$$

The vector $\mathbf{z}_i = (z_{i1}, \dots, z_{i7})$ is retained for modeling and explanations.

D. Quality Control Policy

LD and MIC produce binary flags and soft scores that drive two actions:

- **Hard filter:** exclude records that fail LD or show contradictions in MIC.
- **Soft down-weight** (sensitivity only): assign a weight $w_i \in [0.5, 1]$ to borderline attention or timing cases.

E. Operational, Financial, and Digital Features

All features use information available at time t to avoid leakage.

Operational or Financial (examples): We define the *Outstanding-to-Disbursed Ratio (ODR)* in Eq. (3) and use an indicator for an open *Nonperforming Loan (NPL)* at decision time t ; other examples follow.

$$\text{ODR}_i = \frac{\text{Outstanding}_i(t)}{\text{Disbursed}_i(t)} \quad (3)$$

$$\text{LoanCnt}_i = \# \text{ loans } [t-365, t] \quad (4)$$

$$\text{RecentLoan}_i = \text{days since last loan} \quad (5)$$

$$\text{AvgLoanChg}_i = \Delta_{1y} \overline{\text{Disbursed}} \quad (6)$$

$$\text{OpenNPL}_i = \mathbb{I}\{\text{open NPL at } t\} \quad (7)$$

$$\text{SalaryCnt}_i = \# \text{ salary credits (70/30)} \quad (8)$$

$$\text{AvgSalaryChg}_i = \Delta_{1y} \overline{\text{Salary}} \quad (9)$$

Digital or Demographic: Binary footprint flags (`HasFB`, `HasPhone2`, `HasEmail`) and demographics (`Age`, `Gender`, `EduLevel`). All categoricals include an explicit Missing level.

F. Preprocessing

Numeric missings use median imputation fit on training folds. Heavy-tailed monetary variables are winsorized. Categorical variables use target or mean encoding fit *out-of-fold (OOF)* and applied to validation and test. Psychometric subscales are standardized on training folds before PCS aggregation.

G. Leakage Prevention and Partitions

Data are split 70/15/15 into train, validation, and test. When timestamps permit, splits are out of time; otherwise stratified sampling preserves prevalence. All scalers and encoders are fit on training folds only. The outcome window is defined before feature construction, and records without sufficient observation windows are removed.

H. Modeling and Hyperparameter Search

Baselines include Logistic Regression, *Linear Discriminant Analysis (LDA)*, Decision Tree, Random Forest, *K-Nearest Neighbors (KNN)*, and *Adaptive Boosting (AdaBoost)*. The primary model is LightGBM with class weights and

$$\text{scale_pos_weight} \approx \frac{N_{\text{neg}}}{N_{\text{pos}}}. \quad (10)$$

Hyperparameters are tuned by randomized search with early stopping on validation ROC-AUC and fixed seeds [5].

I. Calibration and Policy Tuning

We report ROC-AUC, PR-AUC, *Kolmogorov-Smirnov (KS)* statistic, Brier score, and *expected calibration error (ECE)*. Policy thresholds are cost based:

$$J(\tau) = c_{\text{FN}} \cdot \text{FN}(\tau) + c_{\text{FP}} \cdot \text{FP}(\tau), \quad \hat{\tau} = \arg \min_{\tau} J(\tau). \quad (11)$$

A dual-threshold policy approves if $p < \tau_{\text{low}}$, declines if $p > \tau_{\text{high}}$, and routes the middle region to manual review subject to a review budget [6], [7].

Algorithm 2 Training, calibration, and thresholding

- 1: Split data into train, validation, and test (chronological if possible).
 - 2: Tune on train and early stop using validation ROC-AUC.
 - 3: Calibrate validation predictions with Platt or isotonic. Select by Brier or log-loss and reliability.
 - 4: Freeze the best model and calibrator, then score the test set.
 - 5: Choose the operating threshold $\hat{\tau} \leftarrow \arg \min_{\tau} (c_{\text{FN}} \text{FN}(\tau) + c_{\text{FP}} \text{FP}(\tau))$, where FN and FP denote false negative and false positive counts; or maximize the *F1-score* (harmonic mean of precision and recall) or F_{β} with $\beta > 1$.
-

J. Explainability and Reason Codes

TreeSHAP (the tree-ensemble variant of SHAP) provides local attributions $\phi_j(i)$ that are mapped to a controlled glossary for compliant adverse-action reasons. Global explanations include feature importance, SHAP summaries, and dependence plots [8], [14].

K. Fairness Diagnostics and Remediation

We evaluate AUC by subgroup (`Gender` and `EduLevel`) and, at $\hat{\tau}$, TPR, FPR, precision (positive predictive value, *PPV*), negative predictive value (*NPV*), and ECE gaps. If gaps exceed governance limits, we consider group-aware calibration, fair thresholding with group-specific τ_g , and feature review [12], [15].

L. Monitoring, Drift, and Governance

We monitor input stability with the *Population Stability Index (PSI)*, PD calibration with reliability slope and intercept, performance with AUC and KS, and explanation stability using cosine similarity of SHAP vectors. Material drift triggers review or recalibration. Artifacts such as data versions, parameters, model hashes, and reason-code glossaries are logged. Psychometric data follow data minimization and explicit consent, and raw responses are segregated from scoring features with restricted access [9], [10].

IV. EXPERIMENTAL SETUP

A. Data and Label

We analyze 59,590 retail applications with `DEFAULT` $\in \{0, 1\}$ over a fixed post-decision horizon (prevalence $\approx 18\%$). Features at decision time t include: (i) seven psychometric subscales aggregated into PCS via Eq. (2); (ii) operational/financial variables (e.g., `ODR`, `LoanCnt`); and (iii) digital/demographic signals. Records failing LD or MIC (Sec. III) are excluded.

B. Partitions and Leakage Control

Data are split 70/15/15 into train/validation/test; when timestamps permit, splits are chronological, else stratified. All statistics (imputers, scalers, encoders, calibrators) are fit on training folds only and applied forward. The outcome window

is fixed before feature construction; records lacking sufficient history are removed.

C. Preprocessing

Numerics: median imputation and winsorization for heavy tails. Categoricals: OOF target/mean encoding with explicit Missing level. Psychometrics: reverse coding, subscale means with completion threshold ρ (e.g., 70%); standardize on training folds and form PCS (Eq. 2).

D. Models and Calibration

Baselines: Logistic Regression, LDA, Decision Tree, Random Forest, KNN, AdaBoost. Primary: LightGBM with class weights and `scale_pos_weight` $\approx N_{\text{neg}}/N_{\text{pos}}$. Tuning: randomized search with early stopping on validation ROC-AUC (fixed seeds). Calibration: Platt or isotonic on validation predictions; selection by Brier/ECE; freeze model+calibrator before test scoring.

E. Policy/Thresholding

Operating threshold $\hat{\tau}$ minimizes the asymmetric cost in Eq. (11) with $c_{\text{FN}}:c_{\text{FP}} = 5:1$. An optional dual-threshold policy defines approve/review/decline bands subject to a review budget.

F. Metrics, Uncertainty, and Fairness

Report ROC-AUC, PR-AUC, KS; calibration via Brier, ECE (10 quantile bins), and reliability slope/intercept. At $\hat{\tau}$: TPR, FPR, PPV, NPV, F1, and $F_{\beta>1}$. Uncertainty uses a stratified paired bootstrap (1,000 replicates, 95% confidence intervals (CIs)). Subgroup diagnostics by Gender and EduLevel cover AUC, error-rate, and calibration gaps; remediation options include group-aware calibration or thresholds.

V. RESULTS AND DISCUSSION

A. Overall Performance

Adding PCS to LightGBM yields consistent gains in ranking and probability quality on the test set. As summarized in Table I, ROC-AUC improves from 0.68 to 0.71, PR-AUC rises from 0.36 to 0.40, and the Brier score drops from 0.167 to 0.161. Figure 1 visualizes the lift across thresholds in both ROC and PR space. In the PR panel, the dashed horizontal line marks the default prevalence (≈ 0.18); the PCS curve remains well above this baseline, which is informative under class imbalance [13]. Improvements over linear baselines indicate that both nonlinearity and the PCS signal contribute to better rank ordering and calibrated PDs.

B. Operating Point and Cost Trade-offs

With the asymmetric cost $c_{\text{FN}}:c_{\text{FP}} = 5:1$ in Eq. (11), the PCS model maintains high specificity while increasing recall (Table II). Holding specificity near 0.90, recall improves from 0.27 to 0.30; precision, NPV, and F1 also increase (from 0.38 to 0.40, 0.90 to 0.91, and 0.31 to 0.34, respectively). Figure 2 (left) shows a clear minimum of total cost near the selected threshold; the confusion matrix (right) makes the error trade-offs explicit via row-normalized rates.

TABLE I
TEST PERFORMANCE. LOWER BRIER IS BETTER.

Model	ROC-AUC	PR-AUC	Brier
Logistic Regression	0.62	0.28	0.184
Linear Discriminant Analysis	0.61	0.27	0.186
Decision Tree	0.60	0.26	0.189
Random Forest	0.66	0.33	0.172
AdaBoost	0.67	0.34	0.170
K-Nearest Neighbors	0.63	0.29	0.183
LightGBM (w/o PCS)	0.68	0.36	0.167
LightGBM (PCS)	0.71	0.40	0.161

TABLE II
THRESHOLDED METRICS AT THE SELECTED OPERATING POINT (TEST SET).

Metric	LightGBM (PCS)	LightGBM (w/o PCS)
Specificity (TNR)	0.90	0.90
Sensitivity (TPR)	0.30	0.27
Precision (PPV)	0.40	0.38
NPV	0.91	0.90
F1	0.34	0.31
Default prevalence	0.18	0.18

TABLE III
ABLATION STUDY (TEST SET).

Variant	ROC-AUC	PR-AUC	Brier
PCS only	0.63	0.30	0.176
No psychological (No PCS)	0.68	0.36	0.167
No digital	0.70	0.39	0.163
No operational/financial	0.62	0.28	0.182
Full (PCS + All)	0.71	0.40	0.161

Portfolio view.: Interpreting the decision rule as “approve when predicted non-default,” the implied decline rate is $\sim 13.6\%$ and the acceptance rate is $\sim 86.4\%$. Among accepted loans, the expected bad rate is $\sim 14.6\%$, a relative reduction of $\approx 19\%$ versus the unconditional 18% prevalence. These quantities are consistent with Table II and Figure 2.

C. Calibration

PCS also improves probability calibration. The reliability curve in Figure 3 lies close to the 45° line with mild tail deviations, aligning with the Brier reduction from 0.167 to 0.161 and low ECE (Sec. IV). Better calibration supports pricing, provisioning, and capital allocation decisions that depend on accurate PDs [6], [7].

D. Ablations and Robustness

Ablation results in Table III confirm that PCS contributes independent, complementary signal. Using PCS alone reaches ROC-AUC 0.63 and Brier 0.176. Removing PCS from the full feature set degrades ROC-AUC/PR-AUC and increases Brier (0.68/0.36/0.167), whereas the full model with PCS performs best (0.71/0.40/0.161). Excluding digital-footprint flags has a smaller effect than removing operational/financial variables, suggesting those remain core predictors.

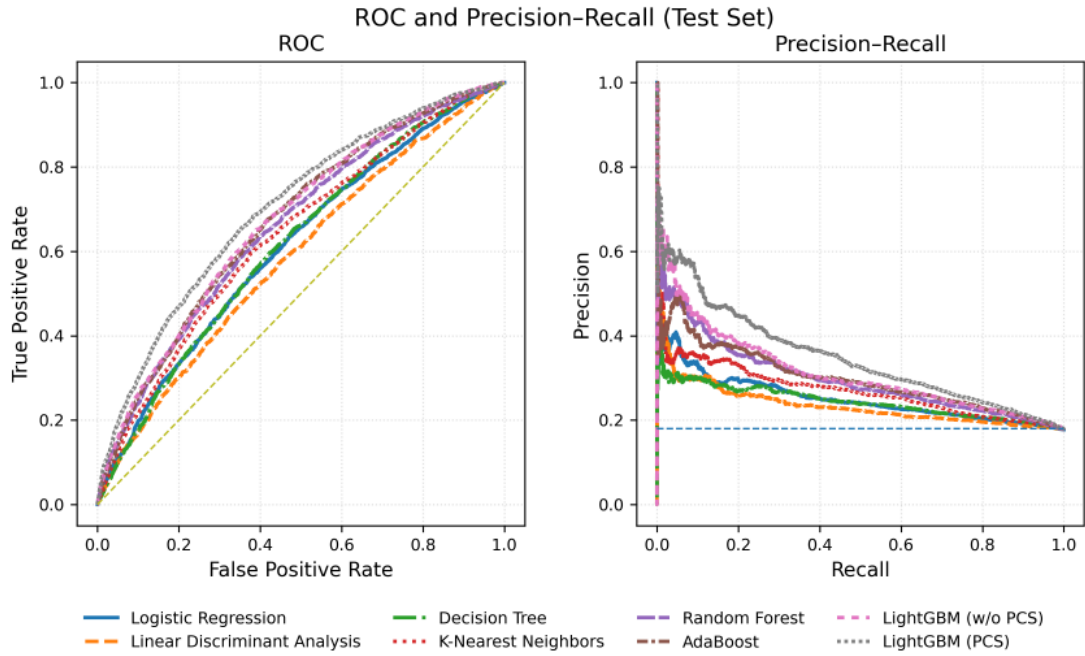


Fig. 1. Ranking performance on the test set: ROC (left) and Precision–Recall (right). The PCS model dominates across thresholds; the dashed PR baseline equals default prevalence (≈ 0.18).

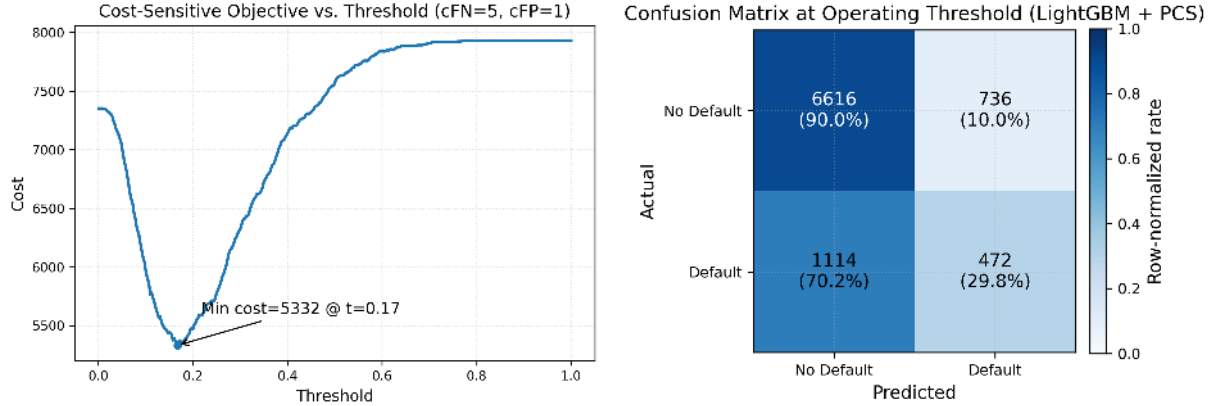


Fig. 2. Operating point visualization. Left: total cost vs. threshold under $c_{FN}:c_{FP} = 5:1$, with the selected threshold marked. Right: row–normalized confusion matrix at that threshold.

E. Explanations and Fairness

Global SHAP patterns (Figure 4) rank *Financial Responsibility (FR)*, *Rule Obedience (RO)*, and *Money Management (MM)* alongside operational features such as ODR and LoanCnt. Higher FR/RO/MM associate with lower risk with diminishing returns, while higher ODR increases risk with flattening at extremes. These findings map directly to reason codes suitable for review and disclosure [8], [14]. Subgroup diagnostics by Gender and EduLevel show small gaps in AUC and thresholded error rates at the chosen operating point. If governance requires tighter parity, group–aware calibration or thresholding can reduce gaps with acceptable efficiency trade–offs [12], [15].

F. Policy Sensitivity

Increasing the FN/FP cost ratio in Eq. (11) shifts the optimal threshold leftward on Figure 2, trading one to two points of specificity for higher recall. When review capacity exists, a dual–threshold policy that routes 5–10% of uncertain cases to manual review can further raise recall with limited precision impact. Under class imbalance, PR curves remain the primary diagnostic for policy selection, to be considered alongside ROC [13].

Takeaway. Across discrimination, calibration, and cost, PCS provides orthogonal, interpretable signal. The selected policy captures more high–risk borrowers at similar specificity, while maintaining explanation quality and acceptable subgroup behavior as supported by Figures 1, 2, 3, and 4.

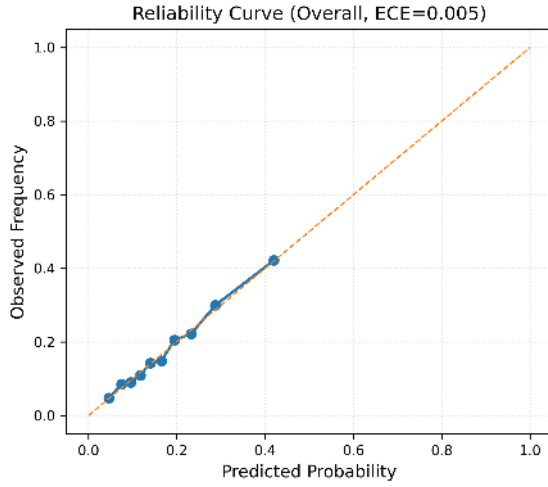


Fig. 3. Reliability curve for LightGBM with PCS on the test set. The dashed line indicates perfect calibration.

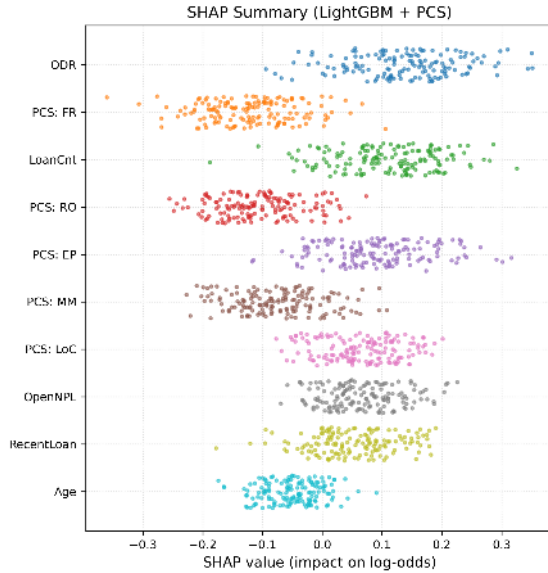


Fig. 4. Global SHAP summary for LightGBM with PCS. PCS subscales (FR, RO, MM) rank alongside key operational features such as ODR and LoanCnt.

VI. CONCLUSION

We introduced PsyFint, a behavior-aware credit scoring framework that augments standard underwriting with a compact PCS. On 59,590 borrowers, adding PCS to a calibrated LightGBM model increased ROC-AUC from 0.68 to 0.71 and reduced the Brier score from 0.167 to 0.161. At a cost-sensitive operating point the model maintained high specificity near 0.90 while increasing recall for high-risk borrowers. PCS subscales, particularly Financial Responsibility, Rule Obedience, and Money Management, ranked alongside repayment-history features in importance, which supports interpretable reason codes and practical deployment. The framework includes quality control, probability calibration, cost-aware policy selection, subgroup diagnostics, and monitoring for drift,

which together address key governance requirements.

Limitations and future work. Self-report bias, external validity across markets, and sensitivity to macro conditions remain open issues. Future work will test measurement invariance, shorten or adapt the instrument, explore group-aware calibration and thresholding, quantify policy uncertainty, and evaluate privacy-preserving collaboration such as federated learning. We also plan stress tests for strategic responses and longitudinal drift to support robust production use.

ACKNOWLEDGMENT

We gratefully acknowledge the research assistants, data specialists, and field coordinators whose contributions were vital to data collection, validation, and the development of Psyfint. We also thank the founders of Optimal NMax LLC for their financial and emotional support, and the company's staff for their practical assistance and valuable advice.

REFERENCES

- [1] L. Thomas, J. Crook, and D. Edelman, *Credit scoring and its applications*. SIAM, 2017.
- [2] C. Strömbäck, T. Lind, K. Skagerlund, D. Västfjäll, and G. Tinghög, "Does self-control predict financial behavior and financial well-being?," *Journal of behavioral and experimental finance*, vol. 14, pp. 30–38, 2017.
- [3] I. Arráiz, M. Bruhn, and R. Stucchi, "Psychometrics as a tool to improve credit information," *The World Bank Economic Review*, vol. 30, no. Supplement_1, pp. S67–S76, 2017.
- [4] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the rise of fintechs: Credit scoring using digital footprints," *The Review of Financial Studies*, vol. 33, no. 7, pp. 2845–2897, 2020.
- [5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.
- [7] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön, "Evaluating model calibration in classification," in *The 22nd international conference on artificial intelligence and statistics*, pp. 3459–3467, PMLR, 2019.
- [8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [9] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, 2015.
- [10] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [11] D. J. Hauser, A. Sunderrajan, M. Natarajan, and N. Schwarz, "Prior exposure to instructional manipulation checks does not attenuate survey context effects driven by satisficing or gricean norms," *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, vol. 10, no. 2, pp. 195–220, 2016.
- [12] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.
- [13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.
- [14] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.