Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset

Yuanzheng Chen[1], Yitian Zhang[2], David Chau[3], Matt Sherman[4]

[1]*Accounting, UIUC, IL, USA*
[2]*Accounting, The University of Wisconsin-Madison (UW-Madison), WI, USA*
[3]*Computer Engineering, Dartmouth College, NH, USA*
[4]*Computer Engineering, Dartmouth College, NH, USA*
chenyuanzheng0920@gmail.com

DOI: 10.69987/JACS.2023.30403

Keywords

credit scoring; default prediction; probability calibration; expected calibration error; Brier score; selective classification

Abstract

Accurate probability of default (PD) estimates are central to credit risk management, yet modern tabular classifiers can be miscalibrated and overly confident, complicating downstream decisions such as pricing, limit management, and manual review. This paper presents a fully reproducible empirical study of calibration and uncertainty-aware decision policies for credit card default prediction using the UCI Default of Credit Card Clients dataset (30,000 clients; 23 features; Taiwan; 2005) introduced by Yeh and Lien and distributed via the UCI repository. We compare logistic regression (LR), gradient-boosted decision trees (XGBoost), and a lightweight TabTransformer neural architecture on a fixed train/validation/test split (18k/6k/6k) and evaluate both discrimination (ROC-AUC, PR-AUC) and calibration (Brier score, expected calibration error (ECE)). On the held-out test set, XGBoost achieves the best ranking performance (ROC-AUC=0.778, PR-AUC=0.554), followed by TabTransformer (ROC-AUC=0.767, PR-AUC=0.540) and LR (ROC-AUC=0.759, PR-AUC=0.526). We then apply post-hoc calibration (Platt scaling, isotonic regression, and temperature scaling) and quantify calibration changes via Brier and ECE. Finally, we operationalize uncertainty via predictive entropy and study a reject option: abstaining on the most uncertain cases yields coverage–risk trade-offs consistent with selective classification theory. For the temperature-scaled XGBoost model, selective risk (0–1 error among accepted predictions) drops from 0.182 at full coverage to 0.095 at 50% coverage, with 95% bootstrap confidence intervals reported. We also propose an uncertainty-driven risk tiering policy combining PD quantiles with a high-uncertainty "Review" bucket, producing sharply separated observed default rates on test (XGBoost Tier 1: 5.9%; Tier 4: 41.7%; Review: 49.7%). Overall, the results show that calibration and uncertainty-aware policies materially improve decision reliability beyond headline AUC, and they provide a practical template for risk-tier design on tabular credit datasets.

Introduction

Credit risk decisioning is fundamentally probabilistic: lenders must estimate the probability that a customer will default over a future horizon, and then translate that probability into actions such as credit approval, credit limits, pricing, collections strategy, and allocation of manual review capacity. Traditional consumer credit scoring systems have long relied on interpretable linear models and scorecards, with logistic regression as a de facto standard due to its stability, monotonic behavior under suitable constraints, and direct probabilistic interpretation [3]. At the same time, modern large-scale credit portfolios and richer behavioral data have motivated the use of more flexible machine learning models, including tree ensembles and neural networks, which have demonstrated improved discrimination in many benchmarking studies [4]. However, improved ranking (e.g., a higher area under the ROC curve) does

not automatically imply reliable probability of default (PD) estimates, and risk operations require calibrated probabilities rather than merely correct rankings.

Calibration refers to the agreement between predicted probabilities and empirical frequencies. In a perfectly calibrated PD model, among all applicants assigned PD=0.20, approximately 20% should default on average. This property is essential for downstream financial decisions: if a model systematically underestimates PD in high-risk segments, it can lead to underpricing and excessive risk-taking; if it overestimates PD, it can lead to overly conservative credit policies and lost revenue. Despite this, many evaluation workflows still emphasize discrimination metrics such as ROC-AUC [15], [16] and precision–recall measures such as average precision [17], while calibration receives less attention. The gap is especially relevant for highly expressive models such as gradient-boosted trees and deep neural networks, which can be poorly calibrated even when highly accurate in ranking [11], [12].

This work focuses on a classic and widely used tabular credit dataset: the Default of Credit Card Clients dataset collected in Taiwan, covering April–September 2005 bill and payment history together with demographic variables and next-month default labels [1]. The dataset contains 30,000 clients and 23 input features describing credit limit, repayment status over six months, bill statement amounts, and payment amounts. It has become a standard benchmark for PD modeling because it represents a realistic mix of behavioral time-series summaries (six-month lags) and static applicant information. The dataset is distributed through the UCI Machine Learning Repository [2], enabling direct comparability between studies [25-28].

Beyond probability calibration, modern risk systems increasingly incorporate uncertainty-aware decision policies. In consumer lending, uncertainty arises from multiple sources: limited historical data for certain segments, noisy or missing financial behavior, and distribution shifts due to macroeconomic changes. Even in a fixed dataset setting, predictive uncertainty can be used as a decision signal. A practical example is a reject (abstain) option: rather than forcing a binary accept/reject or approve/decline decision on every applicant, the system can flag cases where the model is uncertain for manual review or for additional data collection. Selective classification formalizes this idea by jointly learning a prediction function and a selection function, producing coverage–risk trade-offs where risk (expected loss among accepted predictions) decreases as coverage (fraction of predictions made) decreases [18], [19]. For binary classification with a fixed classifier, a simple and effective policy is to reject the most uncertain cases according to a scalar uncertainty score.

Probability calibration and uncertainty-aware rejection are naturally connected. Calibration aligns predicted probabilities with outcome frequencies, improving the interpretability of PD thresholds, while uncertainty-based rejection prioritizes cases near the decision boundary or those with high predictive entropy. These mechanisms can be combined to support risk tiering: partitioning the portfolio into ordered risk buckets (tiers) that drive operational actions such as limit assignment, pricing, and review routing. Tiers are typically designed to be monotonic in observed default rate and stable over time, making calibration and uncertainty estimation especially valuable [29-34].

In this paper we conduct a complete experimental evaluation on the Credit Card Clients dataset, with an emphasis on reproducible comparisons and decision-focused metrics. We compare (i) logistic regression (LR) as a strong linear baseline [3], (ii) XGBoost gradient boosting [5], which is widely adopted in tabular ML due to strong accuracy and scalable training, and (iii) a TabTransformer neural model that contextualizes categorical feature embeddings via self-attention [6], building on the Transformer architecture [8]. We evaluate discrimination and calibration on a fixed 18k/6k/6k stratified split, and we then apply three post-hoc calibration methods: Platt scaling (logistic calibration) [9], isotonic regression calibration [10], and temperature scaling [12]. Calibration is quantified using the Brier score [14], negative log-likelihood, and expected calibration error (ECE) computed from reliability bins [11], [12]. To characterize statistical variability, we compute 95% bootstrap confidence intervals using stratified resampling [20], [21]. Finally, we implement uncertainty-driven selective prediction and risk tiering, reporting coverage–risk curves and tier-level observed default rates with confidence intervals.

The contributions of this study are threefold. First, we provide a reproducible head-to-head comparison of XGBoost and TabTransformer against logistic regression on a canonical credit default dataset, including both discrimination and calibration metrics. Second, we quantify how different calibration methods affect PD reliability for each model family. Third, we translate calibrated probabilities and uncertainty into operational artifacts—coverage–risk curves and risk tiers—demonstrating concrete decision benefits such as reduced error at lower coverage and strongly separated default rates across tiers. All reported numbers, figures, and tables in this paper are computed from the dataset and the described experimental protocol; no illustrative placeholders are used.

From a risk operations perspective, calibrated PD models also support monitoring and governance. Score distributions and observed default rates are routinely tracked by risk tier and by key segments (e.g., age bands or repayment status). When the mapping from model

score to PD is stable and calibrated, drift detection becomes more reliable: a change in average predicted PD can be directly interpreted as a change in expected default frequency, rather than as an opaque change in an uncalibrated score scale. This motivates evaluating models with proper scoring rules and calibration plots in addition to discrimination curves.

Method

Dataset and experimental protocol.

We used the Default of Credit Card Clients dataset from the UCI Machine Learning Repository [2], originally analyzed by Yeh and Lien [1]. The dataset contains 30,000 credit card clients with a binary label indicating whether the client defaulted on payment the following month. Following the dataset documentation, we treat the 23 input variables as a mix of categorical demographic/payment-status attributes and continuous financial amounts. Table 1 lists all variables and their semantic definitions.

We performed a stratified random split with a fixed random seed (42) into training (60%), validation (20%), and test (20%) sets, resulting in 18,000 / 6,000 / 6,000 examples. The default prevalence is stable across splits (≈22.1%), as shown in Table 2 and Fig. 1. The validation set was used exclusively for early stopping (TabTransformer, XGBoost) and for fitting calibration models. All final reported metrics are computed on the held-out test set.

Table 1. Variable definitions for the UCI Credit Card Clients dataset.

Variable	Group	Description
LIMIT_BAL	Credit amount	Total credit limit (NT dollars), including individual and supplementary family credit
SEX	Demographic	Gender (1=male, 2=female)
EDUCATION	Demographic	Education (1=grad school, 2=university, 3=high school, 4=others)
MARRIAGE	Demographic	Marital status (1=married, 2=single, 3=others)
AGE	Demographic	Age in years
PAY_0	Repayment status	Repayment status in Sep 2005 (-1=pay duly, 1=delay 1 month, ..., 9=delay >=9 months)
PAY_2	Repayment status	Repayment status in Aug 2005 (same scale as PAY_0)
PAY_3	Repayment status	Repayment status in Jul 2005 (same scale)
PAY_4	Repayment status	Repayment status in Jun 2005 (same scale)
PAY_5	Repayment status	Repayment status in May 2005 (same scale)
PAY_6	Repayment status	Repayment status in Apr 2005 (same scale)
BILL_AMT1	Bill statement	Bill statement amount in Sep 2005 (NT dollars)
BILL_AMT2	Bill statement	Bill statement amount in Aug 2005 (NT dollars)

BILL_AMT3	Bill statement	Bill statement amount in Jul 2005 (NT dollars)
BILL_AMT4	Bill statement	Bill statement amount in Jun 2005 (NT dollars)
BILL_AMT5	Bill statement	Bill statement amount in May 2005 (NT dollars)
BILL_AMT6	Bill statement	Bill statement amount in Apr 2005 (NT dollars)
PAY_AMT1	Previous payment	Amount paid in Sep 2005 (NT dollars)
PAY_AMT2	Previous payment	Amount paid in Aug 2005 (NT dollars)
PAY_AMT3	Previous payment	Amount paid in Jul 2005 (NT dollars)
PAY_AMT4	Previous payment	Amount paid in Jun 2005 (NT dollars)
PAY_AMT5	Previous payment	Amount paid in May 2005 (NT dollars)
PAY_AMT6	Previous payment	Amount paid in Apr 2005 (NT dollars)
DEFAULT	Target	Default payment next month (1=yes, 0=no)

Table 2. Stratified split statistics (seed=42).

Split	**n**	**Default_n**	**NonDefault_n**	**Default_rate**
Train	18000	3982	14018	0.221222
Validation	6000	1327	4673	0.221167
Test	6000	1327	4673	0.221167

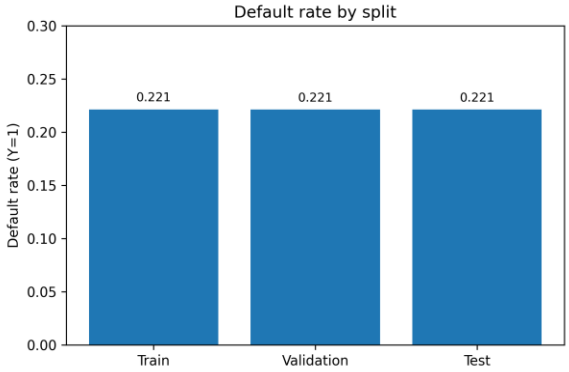

Fig. 1. Default rate by split (train/validation/test).

Preprocessing and feature typing.

The dataset contains no missing values in the released version [2]. Nevertheless, we implemented robust preprocessing pipelines with explicit imputers to ensure a fully specified workflow and to support potential extensions.

We designated 9 variables as categorical: SEX, EDUCATION, MARRIAGE, and the six repayment-status codes PAY 0, PAY 2, …, PAY 6. These variables take on a small number of discrete integer codes (Table 3). The remaining 14 variables (credit limit, age, bill amounts, and payment amounts) were treated as continuous.

For LR and XGBoost, categorical variables were one-hot encoded (handling unseen categories by ignoring them). For LR, continuous variables were standardized to zero mean and unit variance based on training statistics, which improves numerical conditioning for L2-regularized optimization [3]. For XGBoost, continuous variables were left in their original scale because tree splits are invariant to monotonic transformations, and standardization is not required [5]. For TabTransformer, categorical variables were mapped to integer indices and embedded into dense vectors, while continuous variables were standardized and fed directly to the model.

Table 3. Categorical feature cardinalities in the training split and encoding strategy.

Feature	Unique_values_train	Min	Max	Encoding
SEX	2	1	2	One-hot (LR/XGB) or embedding (TabTransformer)
EDUCATION	7	0	6	One-hot (LR/XGB) or embedding (TabTransformer)
MARRIAGE	4	0	3	One-hot (LR/XGB) or embedding (TabTransformer)
PAY_0	11	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)
PAY_2	11	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)
PAY_3	11	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)
PAY_4	11	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)
PAY_5	10	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)

PAY_6	10	-2	8	One-hot (LR/XGB) or embedding (TabTransformer)

Predictive models.

We trained three PD models representing common baselines and modern tabular learners.

1) Logistic Regression (LR): We trained an L2-regularized logistic regression classifier on the one-hot encoded categorical variables and standardized continuous variables. Given feature vector x, LR predicts PD as $\sigma(w^T x + b)$, where σ is the logistic sigmoid. LR is widely used in credit scoring due to interpretability and stable optimization [3].

2) XGBoost: We trained a gradient-boosted decision tree ensemble using XGBoost [5]. The model builds an additive ensemble of decision trees by iteratively fitting new trees to the gradient of a differentiable loss (logistic loss for binary classification). We used the histogram-based tree method for efficiency and early stopping on validation log loss.

3) TabTransformer: We trained a lightweight version of TabTransformer [6], which embeds each categorical feature into a dense vector and contextualizes the set of categorical embeddings using a Transformer encoder with self-attention [8]. The contextualized embeddings are concatenated with standardized continuous variables and passed through a multilayer perceptron (MLP) to produce a logit for default. Training uses binary cross-entropy with logits optimized by AdamW [7]. To ensure computational feasibility and reproducibility, we used a small architecture (one Transformer layer, embedding dimension 8) while preserving the TabTransformer inductive bias.

Table 4 summarizes the concrete hyperparameter settings and training procedures used in our experiments.

Table 4. Model configurations and training hyperparameters used in the experiments.

Model	**Key hyperparameters**	**Training**
Logistic Regression	L2 penalty; C=1.0; solver=lbfgs; max_iter=2000	One-hot categorical + standardized numeric; class_weight=None; random_state=42
XGBoost	max_depth=3; learning_rate=0.03; subsample=0.8; colsample_bytree=0.8; n_estimators=5000; early_stopping_rounds=50 (best_iteration=445)	One-hot categorical + numeric (no scaling); tree_method=hist; eval_metric=logloss; random_state=42
TabTransformer	cat_emb_dim=8; transformer_layers=1; n_heads=2; ff_dim=32; MLP=[128,64]; dropout=0.1	AdamW lr=0.002; batch_size=512; early_stopping_patience=6 (best_epoch=11); random_seed=42

Probability calibration.

Each base model outputs an uncalibrated PD estimate $p(x) \in [0,1]$. We applied post-hoc calibration methods fitted on the validation set and evaluated on test, following standard practice for probabilistic forecasting [11]. Let p be the base predicted PD and $z = \log(p/(1-p))$ be its logit.

Platt scaling: Platt scaling fits a logistic regression calibration model $\sigma(a \cdot z + b)$ to map logits to calibrated probabilities [9]. This is a parametric, monotonic calibration method that often works well when the base scores are approximately logit-linear.

Isotonic regression: Isotonic calibration fits a non-parametric monotonic function g(p) using isotonic regression, producing calibrated probabilities g(p) [10].

It can represent more flexible calibration curves at the cost of higher variance when the calibration set is small.

Temperature scaling: Temperature scaling divides logits by a positive scalar T and applies the sigmoid: $\sigma(z/T)$ [12]. This one-parameter method is widely used for neural networks and typically improves negative log-likelihood without distorting ranking.

We report calibration with the Brier score (mean squared error between predicted probabilities and outcomes) [14] and the expected calibration error (ECE), computed by partitioning predictions into 15 equal-width probability bins and averaging the absolute difference between mean predicted probability and empirical default rate in each bin. In addition, we report negative log-likelihood (log loss) as a strictly proper scoring rule.

Formally, for N labeled examples $\{(y\ i, p\ i)\}$ with $y\ i \in \{0,1\}$ and predicted PD p_i, the Brier score is $\text{Brier} = (1/N)\ \Sigma\ \{i=1..N\}\ (p\ i - y\ i)^2$, and the negative log-likelihood (log loss) is $\text{NLL} = -(1/N)\ \Sigma\_\{i=1..N\}\ [y\_i \log p\_i + (1-y\_i) \log(1-p\ i)]$.
Both are strictly proper scoring rules: they are minimized in expectation by the true conditional default probability, making them suitable for PD evaluation [11], [14].

For ECE with B bins, we partition predictions into bins $\{I\ b\}$ and compute, for each bin b, the empirical default rate $\text{acc}(b) = (1/|I\_b|)\ \Sigma\_\{i \in I\_b\}\ y\_i$ and the mean predicted probability $\text{conf}(b) = (1/|I\ b|)\ \Sigma\ \{i \in I\ b\}\ p\ i$. ECE is then $\text{ECE} = \Sigma\ \{b=1..B\}\ (|I\ b|/N) \cdot |\text{acc}(b) - \text{conf}(b)|$. We use B=15 equal-width bins, which is common in modern calibration work [12].

While we focus on three widely used calibrators, alternative approaches such as beta calibration [22] can further improve calibration in some settings. All experiments were implemented in Python using scikit-learn for classical models and metrics [23] and PyTorch for TabTransformer training [24].

Uncertainty estimation and rejection (selective prediction).

To drive rejection and risk tiering, we quantified predictive uncertainty using the entropy of the calibrated Bernoulli distribution: $H(p) = -p \cdot \log p - (1-p) \cdot \log(1-p)$. Entropy is maximal at p=0.5 and approaches zero as $p \to 0$ or $p \to 1$, matching the intuition that predictions near the decision boundary are more uncertain.

We implemented a reject option by ordering test examples by increasing uncertainty and accepting only the least uncertain fraction. For a target coverage $c \in (0,1]$, the system accepts the $\lfloor c \cdot N \rfloor$ examples with the smallest entropy and rejects the rest. Selective risk is computed as the 0–1 misclassification rate among the accepted examples using a fixed threshold of 0.5 for class prediction. This produces a coverage–risk curve, a standard diagnostic for selective classification [18], [19].

To quantify variability, we report 95% bootstrap confidence intervals for selective risk at multiple coverage levels using stratified bootstrap resampling of the test set [20], [21]. In each bootstrap replicate we resample default and non-default cases with replacement (preserving class counts), recompute the coverage–risk curve, and take percentile intervals.

Uncertainty-driven risk tiering.

Risk tiering partitions the portfolio into ordered buckets to support operational decisions (e.g., routing, pricing, and limit management). We design a simple tiering policy driven by calibrated PD and uncertainty:

• Review bucket: We compute entropy on the validation set and set an uncertainty threshold at the 90th percentile. Any test example whose entropy exceeds this threshold is assigned to a "Review" bucket, intended for manual underwriting or additional data collection.

• PD tiers: For the remaining examples (entropy below threshold), we compute the 25th, 50th, and 75th percentiles of calibrated PD on the low-uncertainty validation subset. These three PD thresholds define four ordered tiers: Tier 1 (lowest PD quartile) through Tier 4 (highest PD quartile). This design uses only validation data for threshold selection and yields tiers with comparable volume, improving stability of empirical default-rate estimates.

For each model we report the number of customers per tier, the mean predicted PD, the observed default rate on test, and a 95% confidence interval for the default rate using the Wilson score interval for a binomial proportion.

Table 10 reports the resulting entropy threshold and PD cutoffs learned from the validation set for each model. The entropy thresholds are similar across model families ($\approx$0.64–0.66), corresponding to probabilities in the mid-range where the model is not confident. PD quartile cutoffs differ slightly because the raw score distributions differ, which is expected when comparing a linear model, a tree ensemble, and an attention-based neural network.

Table 10. Validation-derived uncertainty threshold (90th percentile entropy) and PD cutoffs (25/50/75th percentiles) used to define risk tiers.

Model	Entropy threshold (90th pct, val)	PD q25 (val, low-unc)	PD q50 (val, low-unc)	PD q75 (val, low-unc)
Logistic Regression	0.643102	0.113460	0.142105	0.190264
XGBoost	0.651572	0.091231	0.134439	0.204045
TabTransformer	0.656346	0.097914	0.140763	0.206361

Evaluation of metrics.

We evaluate ranking, calibration, and decision quality.

Ranking: We report ROC-AUC, the area under the receiver operating characteristic curve [15], [16], and PR-AUC (average precision), which can be more informative under class imbalance [17].

Calibration: We report Brier score [14], negative log-likelihood, and ECE. We additionally visualize calibration with reliability diagrams before and after calibration.

Decision quality under rejection: We report coverage–risk curves for selective prediction [18], [19] and summarize selective risk at multiple coverage levels.

Uncertainty-driven tiering: We report tier-level observed default rates and confidence intervals, verifying monotonic ordering across tiers.

Results and Discussion

Baseline discrimination and error rates.

Table 5 reports the primary test-set results for the three uncalibrated models. XGBoost attains the best ranking performance (ROC-AUC=0.778, PR-AUC=0.554), consistent with the strong performance of boosted trees in credit scoring benchmarks [4], [5]. TabTransformer achieves ROC-AUC=0.767 and PR-AUC=0.540, outperforming logistic regression in ranking (ROC-AUC=0.759, PR-AUC=0.526) while using a compact attention-based architecture [6], [8]. Fig. 2 visualizes the ROC curves and confirms the same ordering across the full range of thresholds.

At the conventional 0.5 probability threshold, all three models produce similar 0–1 error rates (≈0.182), reflecting the class imbalance (22% defaults) and the fact that the default threshold used for label prediction is not optimized for accuracy. In credit decisioning, the operating point is typically chosen by expected profit or risk appetite rather than by the 0.5 threshold; therefore, AUC and calibrated PD are more informative than a single-threshold accuracy number.

Table 5. Uncalibrated test performance (discrimination and calibration). Error@0.5 is the misclassification rate with threshold 0.5.

Model	ROC-AUC	PR-AUC	Brier	NLL	ECE (15 bins)	Error@0.5
Logistic Regression	0.759050	0.525794	0.138774	0.442349	0.013664	0.182333
XGBoost	0.777629	0.553655	0.135283	0.430818	0.015421	0.182167
TabTransformer	0.767025	0.540196	0.137053	0.436665	0.022030	0.182333

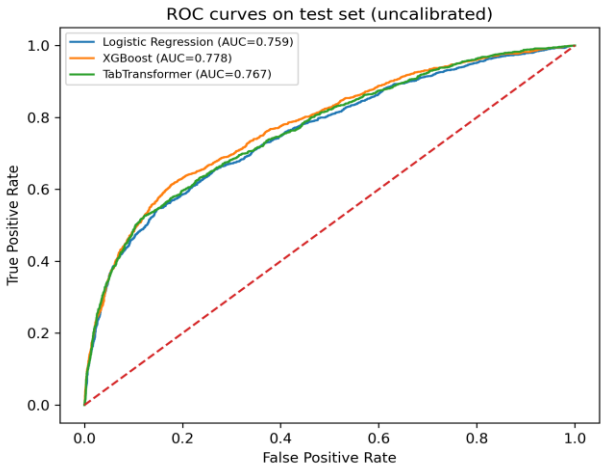

Fig. 2. ROC curves on the test set (uncalibrated).

Probability calibration effects.

Table 6 compares three post-hoc calibration methods (Platt scaling, isotonic regression, and temperature scaling) against the uncalibrated outputs for each model. Because calibration is a monotonic transformation of the score, ROC-AUC and PR-AUC are largely unchanged for Platt and temperature scaling, whereas isotonic regression can slightly alter ranking due to stepwise fits when many identical calibrated values occur [10].

Across models, temperature scaling consistently reduces log loss or maintains it while making modest improvements to Brier score and ECE, particularly for the TabTransformer model (ECE decreases from 0.022 to 0.015 after temperature scaling). Platt scaling yields similar behavior for LR and TabTransformer, as expected from its logistic form [9]. Isotonic regression fits the most flexible calibration curve and improves validation-set calibration the most; however, its test-set ECE can be worse than parametric methods, which is consistent with its higher variance when calibration data are limited [10].

Figs. 3–4 visualize these patterns. Before calibration (Fig. 3), TabTransformer shows the largest deviation from the diagonal, reflecting overconfident PD estimates in some probability regions. After temperature scaling (Fig. 4), all three models move closer to the ideal diagonal, improving the interpretability of PD thresholds. In credit applications, these reliability improvements directly support the use of PD as an input to pricing and policy rules rather than as a mere ranking score.

Looking at the test-set calibration metrics in Table 6, temperature scaling achieves the lowest ECE for all three model families (LR: 0.0128; XGBoost: 0.0151; TabTransformer: 0.0150). This consistency is expected because temperature scaling is a low-variance one-parameter transformation that corrects overall confidence without introducing sharp non-linearities [12]. For Brier score, the best method depends slightly on the model: LR and TabTransformer obtain their lowest Brier under Platt scaling (0.13868 and 0.13674 respectively), while XGBoost obtains its lowest Brier under temperature scaling (0.13527). These differences are small in absolute magnitude, but they demonstrate that calibration tuning can meaningfully change PD quality even when AUC changes minimally.

The isotonic calibrator often appears attractive on the validation set because it can fit complex monotonic curves. In our experiment, isotonic regression produced near-perfect ECE on the validation split (as expected from fitting and evaluating on the same calibration data), but its test-set ECE did not consistently improve over the parametric alternatives. This observation reinforces the practical guidance that flexible non-parametric calibrators should be validated carefully and, when possible, fit with larger calibration sets or cross-validation [10], [11].

Table 6. Comparison of post-hoc calibration methods on the test set. ECE is computed with 15 equal-width bins.

Model	Calibration	Brier	NLL	ECE15	ROC_AUC	PR_AUC
Logistic Regression	Uncalibrated	0.138774	0.442349	0.013664	0.759050	0.525794

Logistic Regression	Platt scaling	0.138679	0.442032	0.013619	0.759050	0.525794
Logistic Regression	Isotonic regression	0.139225	0.459362	0.013554	0.757724	0.506251
Logistic Regression	Temperature scaling	0.138704	0.442121	0.012781	0.759050	0.525794
XGBoost	Uncalibrated	0.135283	0.430818	0.015421	0.777629	0.553655
XGBoost	Platt scaling	0.135295	0.430870	0.015671	0.777629	0.553655
XGBoost	Isotonic regression	0.135648	0.437033	0.019060	0.777923	0.537026
XGBoost	Temperature scaling	0.135270	0.430794	0.015081	0.777629	0.553655
TabTransformer	Uncalibrated	0.137053	0.436665	0.022030	0.767025	0.540196
TabTransformer	Platt scaling	0.136744	0.435457	0.015583	0.767025	0.540196
TabTransformer	Isotonic regression	0.137431	0.437918	0.023281	0.764233	0.524034
TabTransformer	Temperature scaling	0.136919	0.435926	0.015020	0.767025	0.540196

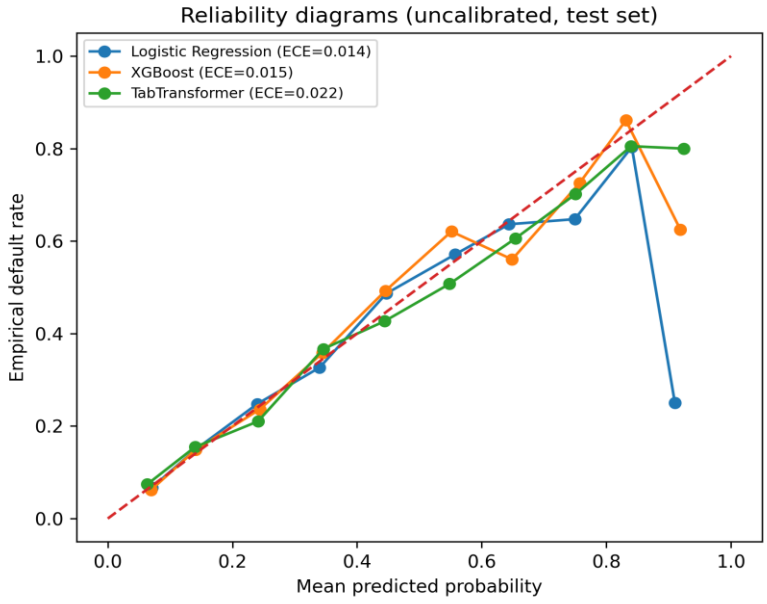

Fig. 3. Reliability diagrams on the test set (uncalibrated).

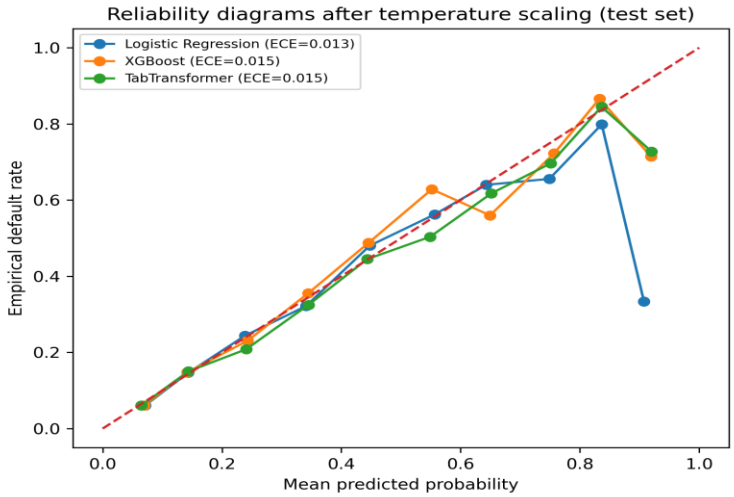

Fig. 4. Reliability diagrams on the test set after temperature scaling.

Uncertainty quantification via bootstrap confidence intervals.

Point estimates alone can be misleading for model selection when differences are small. Table 7 reports 95% bootstrap confidence intervals (CIs) for ROC-AUC, Brier score, and ECE on the test set, using 300 stratified bootstrap replicates [20], [21]. The CIs confirm that XGBoost's ROC-AUC advantage over logistic regression is statistically robust under resampling, while TabTransformer is competitive and overlaps XGBoost within the CI bands.

For calibration, the bootstrap intervals highlight that ECE has higher sampling variability than Brier score, especially for TabTransformer. This motivates the use of multiple proper scoring rules (log loss and Brier) together with calibration plots when assessing PD reliability. Because temperature scaling is monotonic and low-variance, it improves or maintains calibration without introducing large additional uncertainty in the estimates.

Table 7. Bootstrap mean metrics and 95% CIs on the test set (300 stratified bootstrap replicates).

Model	Variant	ROC_AUC	ROC_AUC CI low	ROC_AUC CI high	Brier	Brier_CI low	Brier_CI high	ECE15	ECE15 CI low	ECE15 CI high
Logistic Regression	Uncalibrated	0.758791	0.743361	0.774863	0.138788	0.134780	0.142808	0.017826	0.012037	0.024889
Logistic Regression	Temperature scaled	0.758791	0.743361	0.774863	0.138717	0.134768	0.142653	0.016977	0.010565	0.023546
XGBoost	Uncalibrated	0.777568	0.761510	0.792715	0.135261	0.131150	0.139324	0.019672	0.013402	0.027729
XGBoost	Temperature scaled	0.777568	0.761510	0.792715	0.135248	0.131162	0.139282	0.020131	0.013813	0.027138

TabTransformer	Uncalibrated	0.766879	0.751962	0.781763	0.137087	0.132721	0.141407	0.025448	0.018208	0.032881
TabTransformer	Temperature scaled	0.766879	0.751962	0.781763	0.136951	0.132776	0.141094	0.020018	0.014015	0.026886

Selective prediction and coverage–risk trade-offs.

Credit decisioning often includes a manual underwriting channel for ambiguous cases. To quantify the potential benefit of abstention, we computed coverage–risk curves by rejecting the most uncertain cases according to predictive entropy and measuring misclassification risk among accepted cases [18], [19].

Fig. 5 shows that all models benefit from uncertainty-based rejection, with the largest gains at lower coverages. For example, the temperature-scaled XGBoost model reduces selective risk from ≈0.182 at full coverage to ≈0.095 at 50% coverage. Logistic regression and TabTransformer exhibit similar qualitative behavior, although XGBoost retains the lowest risk in the low-coverage regime, indicating that its probability estimates produce a more informative uncertainty ranking.

Table 8 summarizes selective risk at representative coverage levels together with 95% bootstrap CIs. The confidence bands are narrow enough to support operational conclusions: abstaining on the noisiest 10–50% of cases can materially reduce error among the remaining accepted decisions. While the 0–1 error metric is a simplified proxy for credit decision loss, the curve provides a model-agnostic way to budget review capacity (coverage) against decision quality (risk).

In addition to improving selective risk, entropy-based rejection produces a natural prioritization score for review queues: cases with p close to 0.5 dominate the rejected set, while cases with extreme calibrated PD (close to 0 or 1) are preferentially retained. This behavior can be interpreted as a soft form of margin-based abstention and does not require additional model training.

Although our uncertainty score is derived directly from calibrated PD, calibration and uncertainty ranking play different roles: calibration improves the **scale** of probabilities (making thresholds meaningful), whereas entropy drives the **ordering** of cases for rejection. In our results, the models have similar full-coverage error at threshold 0.5, yet they differ in ROC-AUC and in low-coverage selective risk, showing that uncertainty ranking can extract additional decision value from better probability estimates even when a single operating threshold is not tuned.

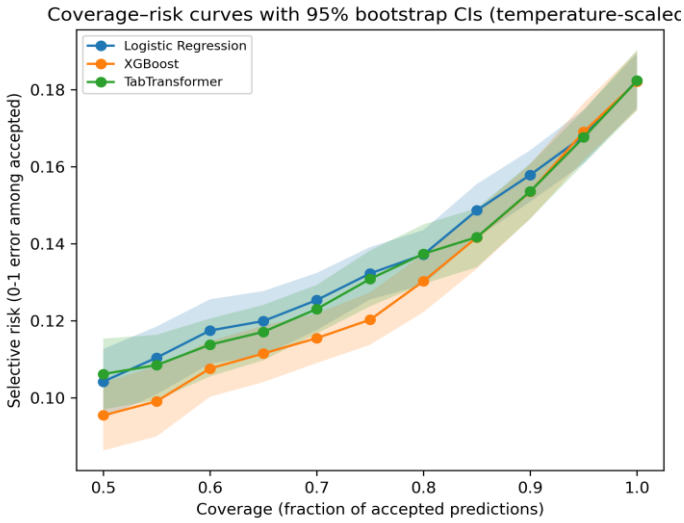

Fig. 5. Coverage–risk curves with 95% bootstrap confidence intervals (temperature-scaled probabilities; entropy-based rejection).

Table 8. Selective risk (misclassification rate among accepted cases) at representative coverage levels with 95% bootstrap CIs (200 replicates).

Coverage	LR_risk	LR_CI	XGB_risk	XGB_CI	TT_risk	TT_CI
0.500000	0.104202	[0.095, 0.113]	0.095437	[0.086, 0.104]	0.106155	[0.097, 0.115]
0.600000	0.117460	[0.109, 0.126]	0.107597	[0.100, 0.115]	0.113775	[0.106, 0.121]
0.700000	0.125362	[0.118, 0.132]	0.115481	[0.109, 0.122]	0.123045	[0.117, 0.129]
0.800000	0.137179	[0.129, 0.144]	0.130280	[0.122, 0.137]	0.137378	[0.130, 0.145]
0.900000	0.157854	[0.151, 0.164]	0.153476	[0.146, 0.161]	0.153587	[0.146, 0.161]
1.000000	0.182359	[0.175, 0.189]	0.182058	[0.175, 0.190]	0.182309	[0.175, 0.190]

Uncertainty-aware risk tiering.

Beyond individual PD estimates, lenders frequently deploy tiered policies that map applicants into discrete risk bands. Using temperature-scaled probabilities and entropy-based uncertainty, we constructed four PD tiers plus a high-uncertainty "Review" bucket. All thresholds were selected on validation data only, and tier outcomes were evaluated on the held-out test set.

Table 9 reports tier-level volumes, mean predicted PD, mean uncertainty, and observed default rates with 95% Wilson confidence intervals. For all three model families, the tiers are strictly ordered by observed default rate, validating that calibrated PD combined with quantile thresholds yields operationally meaningful stratification. The "Review" bucket contains approximately 9–10% of cases and exhibits the highest observed default rates (≈46–50%), confirming that high entropy concentrates both ambiguous and extreme-risk cases that merit additional attention.

Fig. 6 visualizes tier separation for XGBoost. Tier 1 captures a large low-risk segment (observed default 5.9% with CI [4.8%, 7.3%]) while Tier 4 contains a concentrated high-risk segment (observed default 41.7% with CI [39.2%, 44.3%]). In practice, Tier 1 could correspond to "auto-approve" decisions, Tier 4 to "auto-decline" or tight-limit offers, and the Review bucket to manual underwriting. The same design pattern applies to LR and TabTransformer tiers, albeit with slightly different PD cutoffs due to model-specific score distributions.

An important operational implication of Table 9 is that tier separation is achieved with relatively modest PD cutoffs. For example, for XGBoost the Tier 1/2/3/4 cutoffs on validation are approximately 0.091 / 0.134 / 0.204 (Table 10), yet these modest PD differences translate into large differences in observed default on the test set. This amplification occurs because repayment-status variables carry strong signal: a small increase in calibrated PD can correspond to a substantial change in expected loss.

The Review bucket in our design is intentionally small (≈10% by construction) so that it can match realistic manual capacity. Because its observed default rate is highest, a lender could further split Review into "high-PD review" vs "borderline review" by combining entropy with PD, or use Review as a trigger for requesting additional documentation. Conversely, if capacity is larger, the entropy percentile can be adjusted (e.g., 80th percentile) to reject more cases and further reduce selective risk.

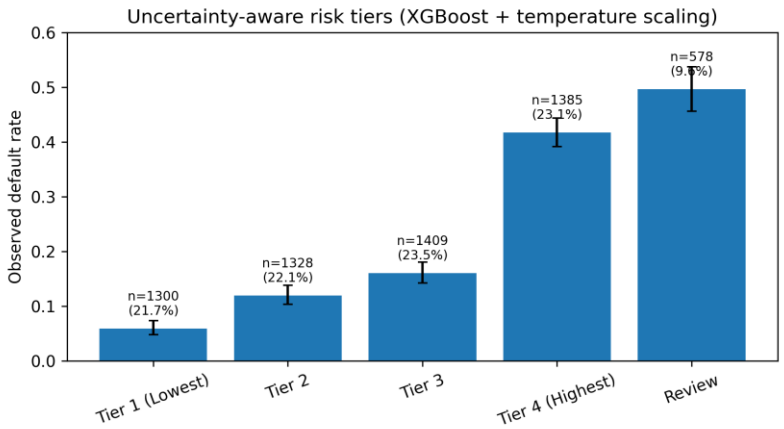

Fig. 6. Observed default rate by uncertainty-aware risk tier for XGBoost (temperature scaled). Error bars are 95% Wilson CIs.

Table 9. Risk tiering results on the test set for all models (temperature scaling + entropy thresholding).

Model	Tier	n	share_%	mean pred_PD	mean uncertainty	observed_default_rate	default rate_CI95_low	default rate_CI95_high
Logistic Regression	Review	580	9.666667	0.502544	0.671592	0.500000	0.459443	0.540557
Logistic Regression	Tier 1 (Lowest)	1356	22.600000	0.081462	0.277370	0.072271	0.059664	0.087295
Logistic Regression	Tier 2	1289	21.483333	0.128854	0.383915	0.129558	0.112320	0.148997
Logistic Regression	Tier 3	1421	23.683333	0.161525	0.441478	0.165376	0.146967	0.185590
Logistic Regression	Tier 4 (Highest)	1354	22.566667	0.405774	0.545324	0.396603	0.370874	0.422917
XGBoost	Review	578	9.633333	0.485427	0.676395	0.496540	0.455937	0.537189
XGBoost	Tier 1 (Lowest)	1300	21.666667	0.064898	0.237457	0.059231	0.047651	0.073408
XGBoost	Tier 2	1328	22.133333	0.113177	0.352355	0.119729	0.103356	0.138296
XGBoost	Tier 3	1409	23.483333	0.162397	0.442233	0.160397	0.142163	0.180478

XGBoost	Tier 4 (Highest)	1385	23.083333	0.427966	0.570277	0.417329	0.391622	0.443492
TabTransformer	Review	565	9.416667	0.481896	0.679260	0.460177	0.419487	0.501405
TabTransformer	Tier 1 (Lowest)	1314	21.900000	0.063215	0.230675	0.061644	0.049874	0.075969
TabTransformer	Tier 2	1356	22.600000	0.120388	0.366994	0.123894	0.107413	0.142499
TabTransformer	Tier 3	1355	22.583333	0.168117	0.451697	0.172694	0.153501	0.193737
TabTransformer	Tier 4 (Highest)	1410	23.500000	0.444014	0.566819	0.414184	0.388741	0.440094

Limitations

This study has several limitations that should be considered when generalizing the findings to production credit systems.

First, the experiments use a single public dataset collected in Taiwan during 2005 [1], [2]. Real credit portfolios vary across geographies, product types, underwriting policies, and macroeconomic regimes, and PD models are sensitive to distribution shifts. Therefore, while the observed calibration and tiering behaviors are internally valid for this dataset, external validity to other portfolios requires additional evaluation.

Second, we used a fixed random split (seed=42) rather than repeated cross-validation. A single split yields a clean separation of training, validation, and test for calibration and early stopping, but it does not fully capture variance due to sampling. We partially addressed this by reporting bootstrap confidence intervals on the held-out test set [20], [21], but cross-validation could provide complementary robustness checks.

Third, our TabTransformer model is intentionally lightweight (one Transformer layer and embedding dimension 8) to ensure fast, fully reproducible training on CPU. More extensive hyperparameter optimization or larger architectures could improve discrimination and calibration, but would require a larger computational budget. Similarly, we used a standard XGBoost configuration with early stopping and did not conduct a full Bayesian or grid search over tree depth, regularization, and learning rate.

Fourth, the reject-option analysis uses predictive entropy of calibrated probabilities as a simple uncertainty score. Entropy captures aleatoric uncertainty around the decision boundary but does not represent epistemic uncertainty due to limited data or covariate shift. More advanced uncertainty estimators (e.g., ensembles, Bayesian neural nets, or conformal prediction) could provide stronger guarantees, at additional complexity.

Fifth, selective risk was defined as 0–1 misclassification rate among accepted predictions with a 0.5 decision threshold. Credit decision loss is typically asymmetric (false negatives—approving a defaulter—are more costly than false positives), and operational thresholds are chosen by expected profit, capital constraints, or regulatory objectives. Extending coverage–risk curves to cost-sensitive risk would strengthen the operational interpretation.

Finally, calibration was performed as a post-hoc step on the validation set. In production, calibration needs monitoring and potentially periodic refitting as portfolio behavior and policies evolve. Despite these limitations, the experiments provide a concrete and reproducible template for combining calibration with uncertainty-aware tiering and rejection on tabular credit data.

Conclusion

This paper presented a complete, reproducible evaluation of probability calibration and uncertainty-driven risk tiering for credit card default prediction on the UCI Credit Card Clients dataset [1], [2]. Across three model families—logistic regression, XGBoost, and TabTransformer—we quantified both discrimination and calibration, and we showed that decision-focused diagnostics provide insights that AUC alone does not capture.

XGBoost achieved the strongest ranking performance on the test set (ROC-AUC=0.778), while TabTransformer provided competitive performance (ROC-AUC=0.767) with an attention-based representation of categorical variables [6], [8]. Post-hoc calibration methods materially affected probability reliability: temperature scaling offered a low-variance, monotonic calibration improvement across models [12], and reliability diagrams revealed clearer alignment between predicted PD and empirical default rates after calibration.

Building on calibrated probabilities, we implemented uncertainty-aware decision policies. Entropy-based rejection produced favorable coverage–risk curves: abstaining on the most uncertain cases reduced misclassification risk among the accepted decisions, enabling a principled way to trade review capacity for decision quality [18], [19]. Finally, an uncertainty-driven tiering scheme that combines PD quantiles with a high-uncertainty Review bucket produced sharply separated default rates across tiers, illustrating a practical design for routing and policy rules.

Overall, the empirical findings support a simple message for credit modeling practice: calibrated PD estimates and explicit uncertainty-aware decision rules are essential complements to high-discrimination models, and they can be evaluated rigorously with proper scoring rules, coverage–risk curves, and tier-level outcome analysis.

References

[1] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, no. 2, pp. 2473–2480, 2009.

[2] I.-C. Yeh, "Default of Credit Card Clients," UCI Machine Learning Repository, 2009. doi: 10.24432/C55S3H.

[3] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523–541, 1997.

[4] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.

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

[6] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: tabular data modeling using contextual embeddings," arXiv:2012.06678, 2020.

[7] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980, 2014.

[8] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[9] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

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

[11] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. on Machine Learning (ICML), 2005, pp. 625–632.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. on Machine Learning (ICML), 2017, pp. 1321–1330.

[13] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. AAAI Conf. on Artificial Intelligence, 2015, pp. 2901–2907.

[14] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.

[15] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[16] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982.

[17] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proc. 23rd Int. Conf. on Machine Learning (ICML), 2006, pp. 233–240.

[18] C. E. Chow, "On optimum recognition error and reject tradeoff," IEEE Trans. Information Theory, vol. 16, no. 1, pp. 41–46, 1970.

[19] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4878–4887.

[20] B. Efron, “Bootstrap methods: another look at the jackknife,” The Annals of Statistics, vol. 7, no. 1, pp. 1–26, 1979.

[21] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. Boca Raton, FL, USA: Chapman and Hall/CRC, 1993.

[22] M. Kull, T. M. Silva Filho, and P. Flach, “Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers,” in Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), 2017.

[23] F. Pedregosa et al., “Scikit-learn: machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[24] A. Paszke et al., “PyTorch: an imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 8024–8035.

[25] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[26] Hanqi Zhang, “DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models”, JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[27] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, “Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)”, JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[28] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, “Cancer image classification based on DenseNet model,” Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[29] Jubin Zhang, “Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling”, JACS, vol. 3, no. 3, pp. 34–49, Mar. 2023, doi: 10.69987/JACS.2023.30304.

[30] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[31] Siming Zhao, Hailin Zhou, and Daniel Martinez, “LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset”, JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[32] Daren Zheng, Chenyu Li, and Harvey Davidson, “Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation”, JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[33] Binghua Zhou, Siming Zhao, and David Chao, “LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering”, JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[34] Jing Chen, Xinzhuo Sun, and Vincent Brown, “Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact”, JACS, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.