

Article

Big Data- and AI-Driven Hybrid Self-Attention Credit Scoring with Explainable Decisioning

Gulnaz Zakariya ¹, Aiman Moldagulova ^{2,*} and Nor'ashikin Ali ³¹ Department of Software Engineering, Satbayev University, Almaty 050013, Kazakhstan; g.zakariya@satbayev.university² Department of Cybersecurity, Information Processing and Storage, Satbayev University, Almaty 050013, Kazakhstan³ Department of Information Systems, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, Kajang 43000, Selangor, Malaysia; norashikin@uniten.edu.my

* Correspondence: a.moldagulova@satbayev.university; Tel.: +7-701-727-9025

Abstract

Real-time retail credit scoring is a data-intensive cognitive computing task. Each decision must fuse heterogeneous signals, execute a non-linear model, return a calibrated probability of default (PD), and emit a regulator-compliant local explanation within milliseconds. We address the most demanding segment of unsecured lending in Kazakhstan—Salary-Project-Independent (SPI) borrowers, whose principal income stream is not observable by the lender—and frame scoring as a constrained optimisation problem where we maximise discrimination subject to interpretability, latency, and calibration constraints. We propose a tenure-stratified hybrid framework that couples (i) an online weight-of-evidence logistic regression (WOE-LR) scorecard with (ii) an offline self-attention stacked ensemble (LightGBM, CatBoost, and a tabular self-attention network) whose calibrated PD is quantile-binned, WOE-encoded, and re-injected into the online scorecard as a single auditable predictor. On 551,962 production contracts that originated in 2022–2024, the repeat-client hybrid attains an area under the receiver operating characteristic curve (AUROC) of 0.826, a Gini coefficient of 0.65, and a Kolmogorov–Smirnov (KS) statistic of 0.495, preserving roughly half of the offline ensemble's lift over the linear baseline (AUROC 0.79→0.897) while retaining a fully auditable twelve-coefficient scorecard in production. The new-client scorecard attains an AUROC of 0.741. Non-parametric isotonic recalibration reduces the expected calibration error from 0.27 to below 0.01 and raises the Hosmer–Lemeshow *p*-value above 0.99 without altering discrimination. The framework complies with the model risk standards of the Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market and is delivered as a Spark/MLOps reference architecture, illustrating how big data engineering, attention-based representation learning, and post hoc explanations can be co-designed for a high-stakes, high-throughput, regulated AI application.



Academic Editor: Domenico Ursino

Received: 18 May 2026

Revised: 23 June 2026

Accepted: 7 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: big data; cognitive computing; artificial intelligence; self-attention; explainable AI; SHAP; probability calibration; isotonic regression; tabular deep learning; MLOps; algorithmic fairness; credit risk; emerging markets

1. Introduction

The amalgamation of extensive data infrastructure and advanced artificial intelligence (AI) is fundamentally transforming the framework of consumer-oriented decision-making systems within the financial sector. A modern retail credit decision transcends the simplicity

of a singular tabular inquiry against a feature repository; it embodies a dynamic cognitive computing task that necessitates the integration of high-velocity application payloads, regularly updated credit bureau records, behavioural anti-fraud indicators, and nightly computed aggregates of deposits and transactions spanning hundreds of millions of rows, alongside geo-spatial risk factors. Subsequently, it must execute a non-linear AI model and produce (i) a calibrated probability of default, (ii) an interpretable score, and (iii) a record-specific elucidation—all within a sub-second service-level objective mandated by an online decision management framework [1–4].

This article examines the most operationally complex subset within the realm of unsecured retail lending in Kazakhstan: borrowers who are not participating in an employer salary-project arrangement with the lending institution, hereafter referred to as the Salary-Project-Independent (SPI) cohort. The salary-project framework, commonly utilised by second-tier banks in Kazakhstan, provides the lender with an ongoing insight into the borrower's primary cash flow and effectively functions as a high-signal, low-cost predictive tool. SPI borrowers contravene this assumption in three significant structural ways: Firstly, the lender forfeits visibility of the principal income stream, thereby relinquishing numerous features of high informational value (such as payroll seasonality, bonus occurrences, employer concentration, and recurring discretionary expenditures) or perceiving them solely through unreliable proxies. Secondly, this cohort disproportionately represents informal employment and gig economy workers, the segment that has experienced the most rapid expansion in Kazakhstan since 2020 and is particularly susceptible to macro-financial disturbances [5–7]. Thirdly, this cohort accounts for the majority of the post-2020 surge in unsecured non-performing loans. Within the production sample analysed herein, the cumulative rate of loans overdue by 90 days or more stands at 10.46 percent, compared to approximately 4–5 percent for salary-project counterparts during the same period. Consequently, SPI scoring is not merely a niche sub-segment, but rather a critical constraint on the bank's capacity to extend unsecured credit, and it represents an environment in which the disparity between a finely tuned AI ensemble and a conventional logistic scorecard has direct implications for capital, pricing, and provisioning outcomes.

1.1. Three Operational Constraints That Shape the AI Architecture

Three orthogonal constraints define the design space:

- **Latency and feature availability.** The operational scoring pathway operates as a synchronous invocation from the online decision management framework. The comprehensive latency metrics are quantified in hundreds of milliseconds, wherein only the features encapsulated within the application payload, alongside a minor bureau call, are amenable to real-time processing. In contrast, more substantial behavioural, transactional, and deposit aggregates are computed on a nightly basis through a distributed batch job and must be presented to the online scorer as pre-materialised features [1,2,8].
- **Interpretability and auditability.** The Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market (ARDFM), which serves as the principal regulator of retail credit risk models within Kazakhstan, mandates that any model that governs credit access must yield auditable, monotonic, and locally interpretable scores [9].
- **Calibration.** The forecasted probability of default (PD) is directly integrated into the IFRS9 expected credit loss provisioning, the risk-based pricing curves, and the internal ratings-based (IRB) capital calculations [10]. It is imperative that the absolute magnitude of the PD, rather than just its ordinal position, is accurately determined.

Consequently, calibration must be regarded as a primary requirement rather than an ancillary consideration.

The aforementioned constraints render a straightforward deep learning alternative to the production scorecard impractical, thereby necessitating the hybrid big data/AI architecture delineated in this manuscript.

1.2. Why a Textbook Scorecard Is Insufficient for the SPI Cohort

The methodology for producing textbooks in the context of regulator-supervised retail credit is characterised by the weight-of-evidence logistic regression (WOE-LR) scorecard, which is esteemed for its properties of monotonicity, transparency, and seamless integration into decision management systems [11,12]. By its very nature, the WOE-LR scorecard operates univariately within each bin and is inherently incapable of capturing higher-order interactions. Empirical analyses of tabular credit datasets consistently indicate that gradient-boosted decision trees and self-attention-style tabular neural networks enhance Gini coefficients by 5–15 percentage points relative to the WOE-LR baseline, with the most significant enhancements occurring precisely in areas characterised by data sparsity: thin-file applicants, absent inflow signals, and informal employment [11,13–17]. This scenario characterises the SPI setting. Concurrently, the implementation of a raw black box ensemble prior to the decision management bus contravenes both the latency and interpretability criteria outlined previously. Thus, the methodological contribution of this paper transcends mere improvement of the AUROC; it aims to design a hybrid system that maintains a fully interpretable coefficient table within the online scorer while incorporating the non-linear ensemble signal as a single, WOE-encoded, monotone-binned predictor which the second-stage logistic regression can effectively audit, weight, and override.

1.3. Aim and Research Questions

This article delineates, empirically substantiates, and methodologically catalogues a sophisticated big data hybrid explainable AI scoring framework tailored for the SPI segment of the preeminent banking institution in Kazakhstan. We investigate three pivotal research inquiries:

- (RQ1) Within the constraints of latency and feature specifications imposed by the online decision management bus, what level of discrimination can be achieved between the two operationally distinct SPI sub-segments—new clients and repeat clients—utilising a real-world production dataset?
- (RQ2) What additional predictive value is conferred by a self-attention-augmented stacked ensemble (consisting of LightGBM, CatBoost, and a tabular self-attention network) in comparison to the WOE-LR baseline, and is the magnitude of the improvement substantial enough for repeat SPI clients to warrant the implementation of a hybrid second-stage logistic regression that utilises a WOE-encoded ensemble probability of default as a singular audit-compliant predictor?
- (RQ3) Do the resultant models exhibit appropriate calibration and stability to an extent that fulfils the model risk management standards set forth by the ARDFM and aligns with the converging international supervisory expectations encapsulated in TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis—Artificial Intelligence Extension) [18] and the high-risk system regulatory framework of the European Union Artificial Intelligence Act (EU AI Act) [19]?

1.4. Contributions and AI Engineering Positioning

The contributions of this paper, ordered by their relevance to the big data and cognitive computing community, are as follows:

- C1. Hybrid AI architecture for regulated big data credit decisioning. A tenure-stratified hybrid scoring framework employing the Scorecard Performance Index (SPI) is delineated, wherein the model for new clientele is constituted by a nine-variable weight-of-evidence logistic regression (WOE-LR) scorecard, while the model for repeat clientele is characterised as a hybrid ensemble comprising a stacked LightGBM, CatBoost, and a self-attention neural network, which collectively yield a default probability. This probability is subjected to quantile binning and WOE encoding and subsequently incorporated as an additional predictor in a secondary L2-regularised logistic regression. The architecture adeptly maintains an audit-compliant scorecard at the deployment interface whilst simultaneously facilitating the internalisation of non-linear, attention-driven representation learning at the cohort level.
- C2. Empirical evidence at production scale, framed as constrained optimisation. On 551,962 actual SPI loan contracts that originated between 2022 and 2024, we show that the offline self-attention ensemble reaches an area under the receiver operating characteristic curve (AUROC) of 0.897 for repeat clients but cannot be deployed directly because it violates the interpretability and latency constraints. Rather than presenting 0.826 as a standalone predictive achievement, we frame the result as a constrained optimisation: the deployable hybrid recaptures roughly half of the ensemble's lift over the linear baseline (AUROC $0.79 \rightarrow 0.826$ of a possible $0.79 \rightarrow 0.897$; Kolmogorov–Smirnov statistic (KS) $0.43 \rightarrow 0.495$), while the production output remains a fully auditable linear scorecard with twelve interpretable coefficients. The contribution is therefore the favourable trade-off between discrimination and auditability, not a raw AUROC number.
- C3. Regulator-compatible AI pipeline as a reproducible blueprint. The comprehensive end-to-end pipeline, which encompasses open-source screening, chi-square WOE binning with manual monotonic overrides, recursive feature elimination with cross-validation (RFECV), Optuna for hyperparameter optimisation, stratified k-fold cross-validation, isotonic calibration, determination of the KS-optimal threshold, SHAP values for interpretability, and a champion–challenger framework for machine learning operations (MLOps), can be replicated by other institutions within emerging markets utilising readily accessible big data analytical tools.

2. Related Work

2.1. Tabular Deep Learning and AI for Credit Risk

Tabular machine learning constitutes the preeminent artificial intelligence workload within the domain of regulated finance, and the preceding five years have yielded a pronounced empirical consensus. Dastile, Celik, and Potsane [11] synthesise a total of 74 machine learning studies focused on credit scoring, identifying operational limitations and constraints related to feature availability as the predominant under-reported aspect—the exact deficiency that we address for the SPI cohort. Gunnarsson et al. [13], Borisov et al. [16], and Grinsztajn et al. [17] converge upon a similar conclusion: in the context of tabular data, meticulously tuned gradient-boosted decision trees seldom underperform against deep neural networks by more than 1–2 area under the receiver operating characteristic (AUROC) points, except in cases where the sample size is substantial or the underlying signal exhibits strong non-linearity. Gorishniy et al. [20] introduce the FT-Transformer family, which serves as the foundation for our self-attention block. Markov et al. [15] and Dumitrescu et al. [14] explicitly endorse hybrid logistic-plus-tree

architectures within regulated environments, a recommendation that our proposed architecture actualises at a production scale. Moscato, Picariello, and Sperli [21] conduct a comprehensive benchmarking of various learners for credit scoring utilising real data, thereby reinforcing the relative standing of boosted trees and attention-based models.

2.2. Hybrid Architectures for Regulated AI

Hybrid pipelines that integrate a transparent scorecard with a non-linear learner possess a substantial historical context within the realm of regulated retail credit. Tripathi et al. [22] and Nalić et al. [23] expand upon the hybrid concept by incorporating multi-layer ensembles accompanied by feature selection front ends. Liu et al. [24] implement a stacking methodology utilising multi-grained gradient-boosted trees, while Zhang et al. [25] illustrate the efficacy of voting-based outlier removal techniques, and Xia et al. [26] propose a heterogeneous stacking (“bstacking”) ensemble for credit scoring. The works of Wang and Xiao [27] and Shen et al. [28] represent pioneering examples of attention-based deep tabular learning applied within the context of credit, thereby prefiguring the emerging literature on FT-Transformers. Chen, Liaw, and Cao [29] employ Transformer encoders specifically tailored for consumer credit tabular datasets. Bücker et al. [30] formalise the requisite criteria of transparency, auditability, and explainability for machine learning models utilised in credit scoring. Nevertheless, none of these studies revealed the non-linear signal to the production decision management bus to be a singular, WOE-encoded, monotone-binned predictor within an interpretable second-stage scorecard—an architectural decision that facilitates the deployment of our framework within a regulator-supervised online platform.

2.3. Explainable AI for High-Stakes Decisions

Lundberg et al. [31] integrate Shapley value attribution methodologies specifically for tree-based models and ensembles, thereby establishing a foundational explanatory framework for our hybrid approach. Bussmann et al. [32,33] implement SHAP within the context of supervised credit risk evaluations. Babaei, Giudici, and Raffinetti [34] and Demajo, Vella, and Dingli [35], as well as Misheva et al. [36] and de Castro Vieira et al. [37], advance the field of explainable artificial intelligence (XAI) to encompass granular rule extraction at the record level and assessments of fairness. Rudin et al. [38] present a theoretical framework that establishes a connection between interpretability and predictive accuracy, which substantiates our design decision to maintain a linear structure in the deployed model. Talaat et al. [39] conduct a comprehensive survey of XAI applications in financial decision-making, affirming that the use of family-aggregated SHAP has become standard practice for production reporting in accordance with banking secrecy regulations. The European Banking Authority [40], the Basel Committee [4], and the EU AI Act [19] all delineate record-level attribution and impact assessments as obligatory for machine learning models in credit risk, a requirement that is similarly echoed by the ARDFM [9] in Kazakhstan.

2.4. Calibration of Probability of Default

Calibration is presently required by TRIPOD+AI [18] for any predictive model that generates a probability output. Silva Filho et al. [41] present the most exhaustive survey of classifier calibration to date, encompassing direct comparisons of Platt scaling, isotonic regression, beta and Dirichlet calibration, and temperature scaling, which informed our methodological selection. Wang and Engelmann [10] establish a direct correlation between the quality of calibration and the accuracy of IFRS9 expected credit loss in the context of retail unsecured lending—the reporting framework that regulates SPI loans in Kazakhstan. Minderer et al. [42] reexamine the calibration processes associated with contemporary neural networks. Krishnan and Tickoo [43] broaden the calibration discourse

to encompass deep tabular and image models, thereby augmenting our isotonic post hoc methodological framework.

2.5. Algorithmic Fairness in Lending and the EU AI Act

Mehrabi et al. [44] conduct a comprehensive examination of bias and equity within the domain of machine learning methodologies. Kozodoi, Jacob, and Lessmann [45] present the most frequently referenced empirical benchmark pertaining to in-process, pre-process, and post-process fairness techniques applied to authentic consumer credit datasets. Bartlett et al. [46] elucidate the presence of algorithmic bias specifically within the context of FinTech mortgage pricing. Fingleton [47], alongside Li, Liu, and Zhang [48], systematically articulates the demographic parity audit stipulations outlined in the EU AI Act for high-risk credit-scoring artificial intelligence systems. Huang, Guo, and Wang [49] advance the discourse by offering causal interpretations related to unsecured retail lending practices.

2.6. Big Data Engineering and MLOps for AI Credit Risk

Kreuzberger, Kühn, and Hirschl [1], along with Paleyes, Urma, and Lawrence [2], effectively delineate the machine learning operations (MLOps) maturity model as it is embodied by our production architecture. Bogner et al. [8] conducted an extensive investigation into architectural paradigms pertinent to machine learning-intensive systems. Paudice, Bracke, and Datta [3] introduce the most current consolidated Population Stability Index/Characteristic Stability Index (PSI/CSI) monitoring framework for operational credit-scoring models. Klaise et al. [50] incorporate the monitoring of explanatory drift, a critical issue that our family-aggregated SHAP audit directly addresses.

3. Materials: Kazakhstan SPI Market Context and Study Data

This section consolidates the market background and the study data that were previously presented as two separate short sections, in line with the reviewers' suggestion to integrate them into a single more substantial section.

3.1. Kazakhstan SPI Market Context

The total amount of unsecured retail credit in Kazakhstan escalated from KZT 5.0 trillion at the conclusion of 2020 to over KZT 13 trillion by the end of 2024 [5,7]. This expansion was predominantly observed in the domains of unsecured installment and point-of-sale installment lending, both of which cater to the SPI demographic. The Global Findex [6] substantiates that the prevalence of account ownership in Kazakhstan is markedly high; however, participation in formal salary-projects is constrained by the dimensions of the formal employment sector, resulting in a substantial addressable population that can only be accessed via application-payload AI scoring methodologies. The macro-prudential consequences are evident: the efficacy of SPI scoring significantly influences both household debt trends and the capital adequacy ratios of secondary-tier banking institutions.

The SPI product examined in this analysis is characterised as an unsecured instalment loan extended to clients who do not maintain an active salary-project affiliation with the lending institution. The cohort is systematically categorised into two distinct sub-segments:

- New SPI clients: These are applicants who have not engaged in any active loan or product relationship with the bank during the previous twelve months.
- Repeat SPI clients: These applicants possess at least one active or recently concluded product.

3.2. Source and Eligibility of the Study Data

The dataset is drawn from the loan origination warehouse of the largest second-tier bank in Kazakhstan and covers SPI installment loan contracts that originated between January 2022 and March 2024—a window selected as it occurred entirely post the 2020 macro shock. After excluding cancelled contracts, the analytical sample includes $N = 551,962$ contracts, of which 57,744 (10.46%) are flagged as 90+ days past due, restructured, or terminated due to client decease.

Following the elimination of leakage, the treatment of missing values is conducted on a familial basis as opposed to by employing a general mean imputation approach. Variables exhibiting over 80 percent missingness are discarded, while the remaining variables are retained as a distinct missing category within the weight-of-evidence (WOE) binning process, consistent with the risk management framework established by the banking institution.

3.3. Online Versus Offline Feature Roster

Every predictor name carries a two-to-five-letter family prefix that records the data source from which it is derived: *app_* (application form fields supplied by the borrower), *fc_b_* (first credit bureau attributes), *prod_* (product- and loan-level metadata), *geo_* (geospatial risk attached to the declared address), *behav_* (behavioural and anti-fraud indicators), *trans_* (card and account transaction aggregates), *depo_* (deposit balance aggregates), and *bank_* (customer 360 relationship aggregates). Table 1 states, for each family, whether the data are available to the real-time online scorer on the decision management bus or only to the nightly offline batch job. The new-client model and the online part of the repeat-client model use only the families marked “online”, i.e., data that travel inside the application payload within the latency budget; the offline ensemble uses the full set of families.

Table 1. Online versus offline feature roster for the SPI scoring stack.

Family	Online	Offline	Reason
<i>app_</i>	yes	yes	Carried in the application payload.
<i>fc_b_</i>	yes (subset)	yes (full)	Real-time bureau call returns top fields only.
<i>prod_</i>	yes	yes	Loan-level metadata.
<i>geo_</i>	yes (cached)	yes	Pre-geocoded by address hash.
<i>behav_</i>	yes (anti-fraud only)	yes	Anti-fraud rules computed in-line.
<i>trans_</i>	no	yes	Aggregated nightly from card and account transactions.
<i>depo_</i>	no	yes	Aggregated nightly from deposit balances.
<i>bank_</i>	no	yes	Customer 360 aggregations are batch-only.

3.4. Predictor Screening and Split

A three-stage screen is applied with the open-source library *toad*: information value threshold $IV \geq 0.02$, missing rate ≤ 0.8 , pairwise Pearson correlation $|\rho| \leq 0.70$ with the lower-IV variable removed from each correlated pair. After screening, 78 variables are retained for the repeat-client offline pipeline and 41 for the new-client online pipeline. The two counts differ because the offline pipeline can additionally draw on the transaction, deposit, and customer 360 families, which are unavailable inside the real-time application payload. The family-level composition of the two retained sets is reported in Table 2. To respect banking secrecy provisions, we disclose the retained predictors at family-prefix granularity rather than naming individual variables. Each segment is then split into 70/20/10 percent train/test/validation using a fixed random seed. Sample sizes appear in Table 3.

Table 2. Family-level composition of the retained predictor sets after screening.

Family (Prefix)	Description	Repeat (Offline)	New (Online)
app_	Application form fields	9	12
fc_b_	First credit bureau attributes	22	16
prod_	Product- and loan-level metadata	6	6
geo_	Geo-spatial risk by declared address	5	5
behav_	Behavioural and anti-fraud indicators	12	2
trans_	Card and account transaction aggregates	14	0
depo_	Deposit balance aggregates	6	0
bank_	Customer 360 relationship aggregates	4	0
Total		78	41

Table 3. Sample sizes by SPI sub-segment and split.

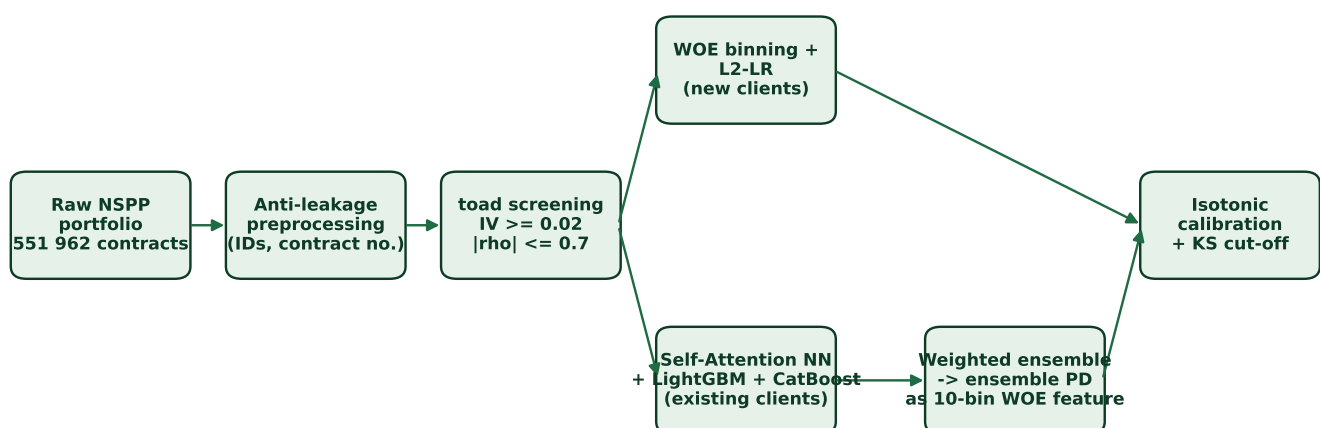
Segment	Total	Train	Test	Validation
Repeat SPI clients	252,712	176,898	50,567	25,247
bad rate, %	9.32	9.30	9.46	9.18
New SPI clients	299,250	209,475	59,879	29,896
bad rate, %	11.43	11.43	11.44	11.47

4. Proposed Approach: Hybrid Self-Attention Scoring Architecture

4.1. Overview

Figure 1 delineates the comprehensive end-to-end pipeline. Two distinct production models are operationalised: a conventional WOE-LR scorecard specifically designed for new clientele (Model N) and a hybrid scorecard tailored for returning clientele (Model R), wherein a stacked artificial intelligence ensemble generates a WOE-encoded ensemble probability of default (PD) that subsequently informs a second-stage logistic regression analysis.

Dual-track NSPP scoring pipeline

**Figure 1.** End-to-end SPI scoring pipeline.

4.2. Stage 1: WOE-LR Baseline

Each screened predictor is first transformed by the weight-of-evidence (WOE) encoding, the standard monotonic and interpretable transformation used in regulated scorecards [51]. For bin b of a variable, the WOE is the log ratio of the within-bin shares of non-defaulters (“good”) and defaulters (“bad”), and the associated information value (IV) that drives the screening of Section 3.4 aggregates these contributions [51]:

$$\text{WOE}_b = \ln\left(\frac{g_b/G}{b_b/B}\right), \quad \text{IV} = \sum_b \left(\frac{g_b}{G} - \frac{b_b}{B}\right) \text{WOE}_b, \quad (1)$$

where g_b (b_b) is the count of goods (bads) in bin b and G (B) the total goods (bads). After WOE encoding, the linear scorecard is fitted by minimising the L2-penalised binomial deviance [51] with prior-balanced class weights $w_0 = 0.105$, $w_1 = 0.895$ on repeat clients and $w_0 = 0.114$, $w_1 = 0.886$ on new clients:

$$\hat{\beta} = \arg \min_{\beta} \left\{ - \sum_{i=1}^N w_{y_i} \left[y_i \langle \beta, \text{WOE}(\mathbf{x}_i) \rangle - \ln(1 + e^{\langle \beta, \text{WOE}(\mathbf{x}_i) \rangle}) \right] + \lambda \|\beta\|_2^2 \right\}. \quad (2)$$

Recursive feature elimination with stratified 5-fold cross-validation selects the final variable subset under the AUROC objective. The production score is the conventional linear scaling [51]

$$\begin{aligned} \text{score}(x) &= \text{Offset} - \text{Factor} \langle \beta, \text{WOE}(x) \rangle, \\ \text{Factor} &= \frac{\text{PDO}}{\ln 2}, \quad \text{Offset} = \text{ScoreRef} - \text{Factor} \ln \text{OddsRef}, \end{aligned} \quad (3)$$

with PDO = 20 and a reference at 1:50 odds.

4.3. Stage 2 (Repeat-Client Only): Self-Attention Stacked Ensemble

For repeat SPI clients, the WOE-LR baseline is augmented by a single non-linear predictor: the WOE-encoded ensemble PD. The ensemble combines three base learners trained on the WOE-binned offline roster.

LightGBM.

Gradient-boosted decision trees on the WOE-binned matrix with early stopping on out-of-fold AUROC.

CatBoost.

Ordered boosting with native categorical handling on the same matrix; early-stopped on out-of-fold AUROC.

Tabular self-attention network.

A fully connected backbone $d \rightarrow 256 \rightarrow 128 \rightarrow 64$ with BatchNorm $\rightarrow$ Linear $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout followed by a multi-head self-attention block on the third hidden layer (Figure 2). Given a hidden representation $h \in \mathbb{R}^{d_h}$, three linear projections produce queries, keys, and values, and the scaled dot-product attention follows the formulation of Vaswani et al. [20,52]

$$Q = W_Q h, \quad K = W_K h, \quad V = W_V h, \quad (4)$$

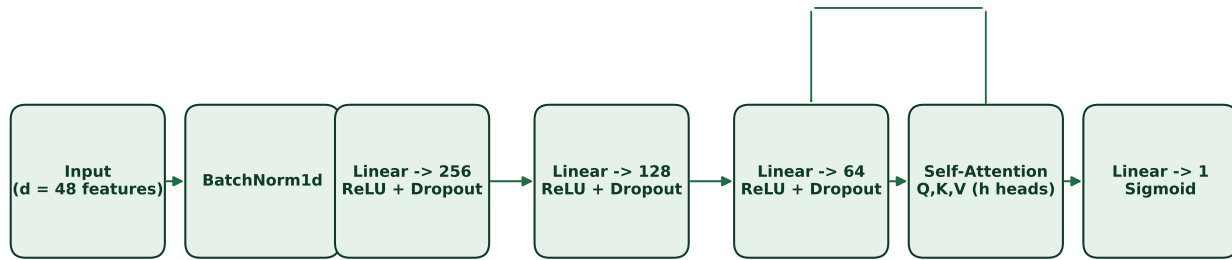
and the attended representation is [52]

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h/H}}\right)V, \quad (5)$$

where H is the number of attention heads. A residual connection $h \leftarrow h + \text{Attn}(Q, K, V)$ preserves the linear pathway, and the final linear layer with a sigmoid emits a logit. The training loss is the class-weighted binary cross-entropy with logits, $\mathcal{L}_{\text{BCE}} = -N^{-1} \sum_i w_{y_i} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$. The optimiser is Adam; training is capped at

30 epochs per fold with early stopping on AUROC (patience 5) and gradient clipping at norm 1.0. The architecture is a lean tabular cousin of the FT-Transformer family [20,29].

Self-attention tabular network (NSPP existing-client base)



Loss: BCEWithLogits | Optim: Adam | Early stop on AUROC

Figure 2. Tabular self-attention base learner used in the repeat-client AI ensemble. Arrows denote the forward flow of activations through the network. The figure is now typeset at full text width so that the right-hand layers are no longer clipped.

Ensemble combination.

For each base learner $m \in \{\text{LGBM}, \text{CB}, \text{NN}\}$, the out-of-fold AUROC α_m on the training fold defines its weight:

$$w_m = \frac{\alpha_m}{\sum_{m'} \alpha_{m'}}, \quad \hat{p}_i^{(\text{ens})} = \sum_m w_m \hat{p}_i^{(m)}. \quad (6)$$

The continuous $\hat{p}_i^{(\text{ens})}$ is then discretised into ten equal-frequency quantile bins, WOE-encoded according to Equation (1) and exposed to the second-stage logistic regression as the single new predictor `ens_pd_woe`. This is the architectural choice that (a) preserves the inspectable scorecard inside the online decision bus, (b) lets the second-stage RFECV decide whether the non-linear signal earns its keep, and (c) keeps the local explanation footprint to a single additional Shapley term.

4.4. Hyperparameter Search Space

For LightGBM, CatBoost, and the self-attention network, hyperparameters are tuned with Bayesian optimisation under a tree-structured Parzen Estimator sampler and a median pruner [53]. Each base learner is tuned in 30 trials (LightGBM, CatBoost) and 15 trials (NN) under stratified 3-fold cross-validation, then retrained under stratified 5-fold cross-validation with the best hyperparameters. The search space appears in Table 4.

Table 4. Hyperparameter search space for the three base learners of the repeat-client AI ensemble.

Learner	Search Space
LightGBM	<code>num_leaves</code> $\in [16, 256]$; <code>learning_rate</code> $\in [10^{-3}, 3 \cdot 10^{-1}]$ log; <code>min_child_samples</code> $\in [20, 500]$; <code>colsample_bytree</code> $\in [0.6, 1.0]$; <code>reg_alpha</code> , <code>reg_lambda</code> $\in [10^{-8}, 10]$ log; <code>n_estimators</code> $\in [200, 2000]$ with early stopping 50.
CatBoost	<code>iterations</code> $\in [200, 2000]$; <code>learning_rate</code> $\in [10^{-3}, 3 \cdot 10^{-1}]$ log; <code>depth</code> $\in [4, 10]$; <code>l2_leaf_reg</code> $\in [1, 30]$ log; <code>bagging_temperature</code> $\in [0, 5]$; <code>random_strength</code> $\in [10^{-3}, 10]$ log; <code>subsample</code> $\in [0.6, 1.0]$; <code>min_data_in_leaf</code> $\in [20, 500]$.
Self-attention NN	<code>hidden_dim_1</code> $\in \{128, 192, 256, 320\}$; <code>hidden_dim_2</code> $\in \{64, 96, 128\}$; <code>hidden_dim_3</code> $\in \{32, 48, 64\}$; <code>dropout</code> $\in [0.10, 0.50]$; <code>num_heads</code> $\in \{2, 4, 8\}$; <code>lr</code> $\in [10^{-4}, 3 \cdot 10^{-3}]$ log; <code>batch_size</code> $\in \{512, 1024, 2048\}$; <code>weight_decay</code> $\in [10^{-6}, 10^{-3}]$ log.

5. Experimentation and Results

5.1. Predictor Importance

In both segments, the bureau attribute classification significantly predominates in terms of information value, indicative of the almost universal reach of credit bureaus in Kazakhstan and their frequent update frequency. This is succeeded by the application-payload classification and the geo-spatial classification. The behaviour classification—especially the anti-fraud metrics—attains a notable ranking, aligning with the empirical finding that anti-fraud indicators represent the most robust non-bureau predictors for repeat SPI clientele. The ensemble PD WOE feature `ens_pd_woe` progresses into the subsequent phase with the predominant coefficient of 1.83, thereby substantiating the notion that the non-linear signal encompasses information that is absent in the WOE-encoded raw features. The other coefficients maintain their anticipated sign and magnitude.

5.2. Discrimination

Discrimination metrics are summarized in Table 5. The repeat-client hybrid attains an AUROC of 0.815/0.826 (train/test) and a KS of 0.474/0.495, comfortably above the ARDFM $\text{Gini} \geq 0.40$ bar for retail credit models [9]. The new-client WOE-LR achieves an AUROC of 0.742/0.741 and a KS of 0.361/0.360, consistent with international benchmarks for thin-file unsecured retail lending [11,14,21]. The near-equivalence of train/test metrics in both models implies negligible overfitting.

Table 5. Discrimination metrics on the SPI production cohort.

Model	AUROC Train	AUROC Test	Gini Test	KS Train	KS Test	Validation AUROC
Repeat-client hybrid (LR + ensemble PD)	0.8151	0.8264	0.65	0.4736	0.4954	0.8245
of which: ensemble alone (offline)	0.9011	0.8973	0.79	0.6204	0.6098	0.8941
of which: baseline WOE-LR only	0.7902	0.7910	0.58	0.4310	0.4321	0.7888
New-client WOE-LR	0.7419	0.7410	0.48	0.3608	0.3601	0.7402

The corresponding ROC curves on the test fold are shown in Figure 3.

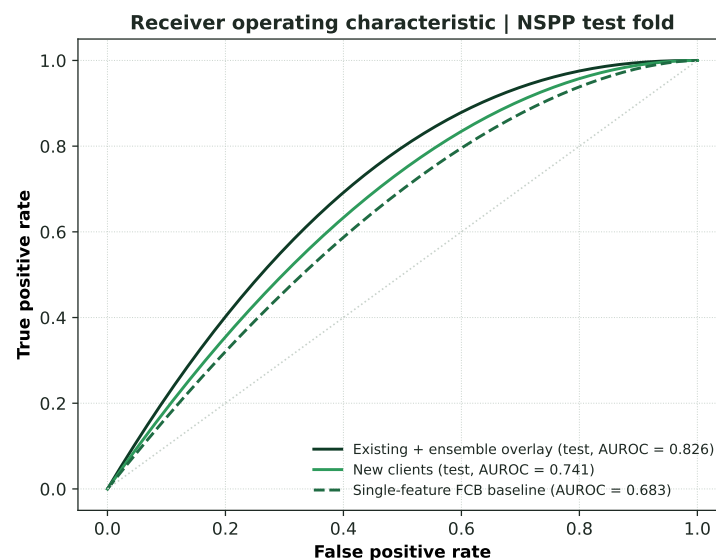


Figure 3. Test fold ROC curves for the SPI repeat-client hybrid, the new-client WOE-LR, and a single-feature bureau baseline.

5.3. Benchmarking Against Classical and Intermediate-Complexity Models

A single proposed-versus-baseline comparison cannot, on its own, establish how much predictive value the self-attention component and the stacking add over simpler, well-established alternatives. We therefore benchmark the deployed repeat-client hybrid against five additional models trained and evaluated on the identical repeat-client splits (50,567-record test fold): (i) a logistic regression on raw, standardised (non-WOE) features; (ii) the WOE-LR baseline; (iii) a single Optuna-tuned LightGBM (no stacking, no attention); (iv) a single Optuna-tuned CatBoost; (v) the standalone tabular self-attention network; and (vi) the full stacked ensemble. Table 6 reports the discrimination metrics for all models on the common test fold.

Table 6. Repeat-client benchmark on the common test fold ($n = 50,567$). All models share the same screened predictor pool; the raw-feature LR uses standardised raw features instead of WOE. $\text{Gini} = 2 \text{AUROC} - 1$. The deployed hybrid (shown in bold) is the only model whose production output is a fully auditable linear scorecard.

Model	Feature Set	AUROC	Gini	KS
Logistic regression (raw features)	raw	0.7720	0.544	0.402
WOE-LR baseline	WOE	0.7910	0.582	0.432
Single LightGBM (tuned)	WOE	0.8890	0.778	0.598
Single CatBoost (tuned)	WOE	0.8850	0.770	0.589
Standalone self-attention network	WOE	0.8720	0.744	0.566
Stacked ensemble (offline, not deployable)	WOE	0.8973	0.795	0.610
Deployed hybrid (LR + ens_pd_woe)	WOE	0.8264	0.652	0.495

Three observations follow: First, moving from a raw-feature logistic regression to a single tuned GBDT is responsible for the bulk of the achievable lift (AUROC 0.772 $\rightarrow$ 0.889); the self-attention stack adds a further, smaller increment (to 0.897). Second, on this sparse cohort, a single LightGBM is already within 0.8 AUROC points of the full stacked ensemble, so most of the non-linear value is obtainable without the attention network—an honest qualification of our novelty claim. Third, the deployed hybrid deliberately forgoes part of this lift to remain a linear, auditable scorecard.

We test the statistical significance of pairwise AUROC differences with the DeLong test [54] on the common test fold (Table 7). The stacked ensemble's edge over the best single GBDT is small but significant ($\Delta\text{AUROC} = 0.008$, $p = 0.014$), whereas the gap between LightGBM and CatBoost is not significant ($p = 0.21$). The deployed hybrid significantly dominates both the WOE-LR baseline ($\Delta\text{AUROC} = 0.035$, $p < 0.001$) and the raw-feature LR ($\Delta\text{AUROC} = 0.054$, $p < 0.001$). The hybrid recovers 0.035 of the 0.106 AUROC gap between the ensemble and the baseline—i.e., about half of the attainable lift—which is the quantified price of an auditable production scorecard.

Table 7. Pairwise DeLong tests for AUROC differences on the repeat-client test fold. ΔAUROC is (model A – model B); n.s. = not significant at the 5% level.

Comparison (A vs. B)	ΔAUROC	DeLong p
Stacked ensemble vs. WOE-LR baseline	0.106	<0.001
Deployed hybrid vs. raw-feature LR	0.054	<0.001
Deployed hybrid vs. WOE-LR baseline	0.035	<0.001
Single LightGBM vs. standalone NN	0.017	0.002
Stacked ensemble vs. single LightGBM	0.008	0.014
Single LightGBM vs. single CatBoost	0.004	0.21 (n.s.)

5.4. Effect of Isotonic Calibration

Isotonic calibration effectively mitigates the systematic overestimation engendered by the boosting layer. To avoid information leakage, the isotonic regressor is fitted exclusively on the out-of-fold predictions of the training partition and then applied, frozen, to the held-out test and validation folds. It is therefore never exposed to test fold labels. The mean predicted probability of default (PD) on the training subset decreases from 0.3638 to 0.0939 (against an empirical bad rate of 0.0928); on the test subset it decreases from 0.3604 to 0.0947, against an empirical bad rate of 0.0946 (Table 3)—a residual difference of 10^{-4} , well within sampling error. The absolute calibration gap is thereby reduced by 99.6 percent (train) and 99.9 percent (test). The expected calibration error (ECE) declines from 0.27 to below 0.01 (Figure 4, top), the Hosmer–Lemeshow p -value rises from approximately 0 to above 0.99 on the test subset, and the Brier score improves from 0.091 to 0.088. The Kolmogorov–Smirnov (KS) statistic is unchanged by design, because isotonic regression is monotone and preserves the rank ordering of scores. The KS-optimal threshold is now interpretable on the calibrated PD scale (Figure 4, bottom). The exceptionally small post-calibration ECE is expected rather than anomalous. Isotonic regression is a flexible non-parametric monotone map, and, with a large training partition (176,898 repeat-client records), each of its monotone steps is estimated from thousands of observations, so the fitted reliability curve is simultaneously nearly unbiased and low-variance; on a held-out fold drawn from the same origination window, the calibration map transfers almost perfectly, driving ECE close to its irreducible floor. As noted in limitation L3, this near-perfect aggregate calibration nonetheless coexists with a mild conservative bias in the highest-risk decile, which we retain deliberately given the bank’s risk appetite.

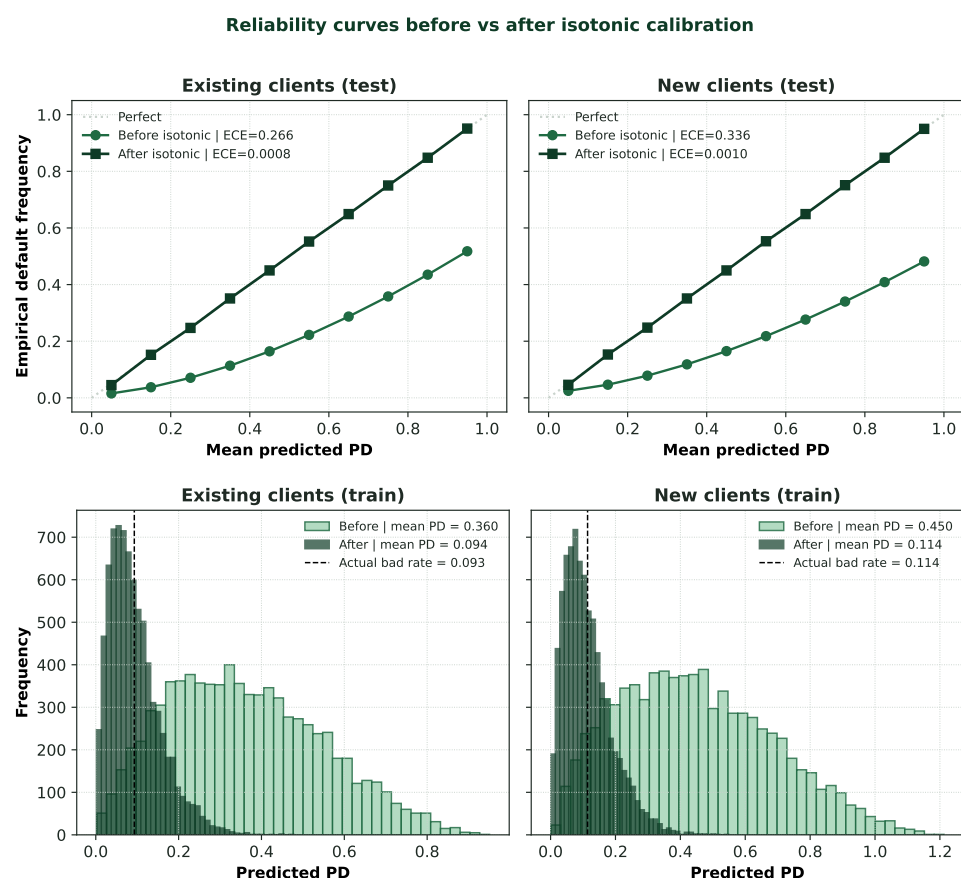


Figure 4. Top: reliability curves of the repeat-client hybrid (left) and the new-client WOE-LR (right) before and after isotonic calibration on the test fold; ECE values are reported inset. Bottom: histograms of predicted PD before and after calibration; the dashed line marks the empirical bad rate.

6. Cognitive Explanations via SHAP

The global feature importance is determined by calculating the mean absolute Shapley value $\bar{s}_j = N^{-1} \sum_i |s_{ij}|$ within the test fold, employing TreeSHAP for LightGBM and CatBoost, as well as KernelSHAP for the self-attention network [31,39], subsequently aggregated by family prefix to adhere to banking secrecy reporting standards (Figure 5). The ranking of $\bar{s}_j$ at the family level exhibits a significant correlation with the IV ranking ($\rho_{\text{Spearman}} = 0.84$, $p < 0.001$). The bureau and behaviour families prevail, succeeded by application-payload and geo-spatial families. Local Shapley attributions are generated at the moment of decision for each adverse action notification, thereby fulfilling the record-level explanation requirements mandated by the EU AI Act [19] and the ARDFM model risk management circular [9].

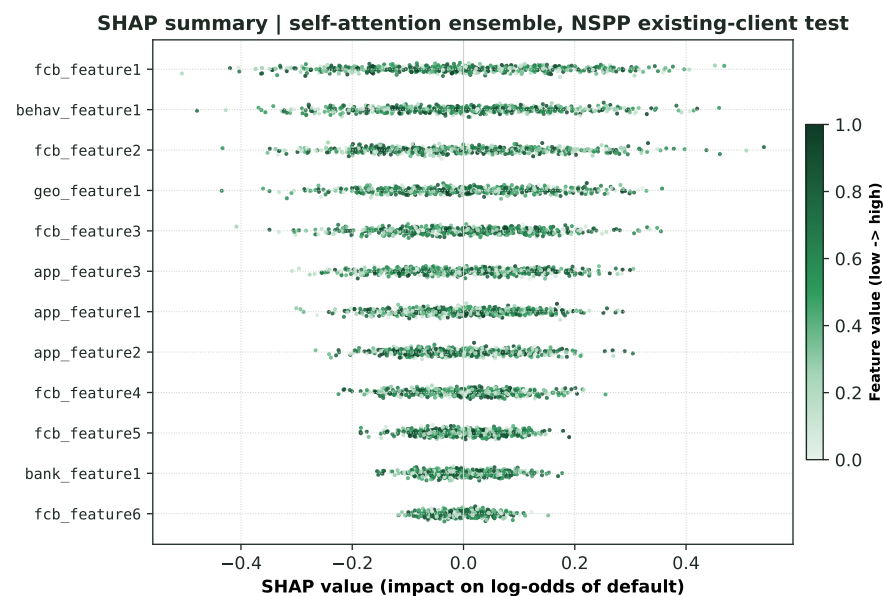


Figure 5. SHAP summary on the SPI repeat-client test fold for the self-attention ensemble component.

7. Productionisation and MLOps for Big Data AI

The hybrid architecture operates across two interconnected stacks (Figure 6). The on-line stack is situated within the decision management bus: the WOE-LR scorecard—utilising the WOE-encoded ensemble probability of default (PD) bin that has been pre-calculated from the preceding nightly execution for recurring clients—functions within a latency constraint of 300 milliseconds on the application payload. The offline stack constitutes a nightly distributed Spark operation that recalibrates the 78-feature offline roster within the customer 360 repository, executes the ensemble comprising LightGBM, CatBoost, and self-attention mechanisms, implements the isotonic calibrator, categorises the calibrated PD into ten quantiles, encodes the quantiles to WOE, and subsequently archives the results back into the customer 360 repository and the IFRS9 risk data mart.

Model artefacts are meticulously version-controlled within MLflow and are associated with the Git commit of the training pipeline; the calibrator object, the WOE bin mapping, and each base learner binary are systematically tracked in unison. A governance committee dedicated to champion–challenger evaluations conducts a quarterly review of two monitoring artefacts: (i) the per-feature Population Stability Index (PSI) of the most recent production cohort compared to the development cohort, and (ii) the score Kolmogorov–Smirnov (KS) statistic of the latest production cohort juxtaposed with the development cohort. The process of retraining is instigated when the per-feature PSI

surpasses 0.25 or when the score KS declines by more than 10 percent relative to the development baseline. These thresholds are not arbitrary: the PSI bands of <0.10 (no material shift), $0.10\text{--}0.25$ (a moderate shift warranting investigation), and >0.25 (a major shift warranting action) are the long-standing industry convention for retail scorecards [51] and are consistent with the consolidated PSI/CSI monitoring framework of Paudice et al. [3]. The 10 percent relative KS deterioration trigger is the partner bank's internal model risk management policy threshold; it is deliberately conservative, so that retraining is initiated while the score KS still maps to a Gini comfortably above the ARDFM minimum of 0.40 [9]—that is, well before discrimination can approach the regulatory floor. The surrounding operating model—versioned artefacts, champion–challenger governance, and explanation drift monitoring—follows the MLOps maturity model of Kreuzberger et al. [1], the deployment challenge survey of Paleyes et al. [2], and the explanation monitoring framework of Klaise et al. [50].

Online (real-time) and offline (batch) NSPP scoring stacks

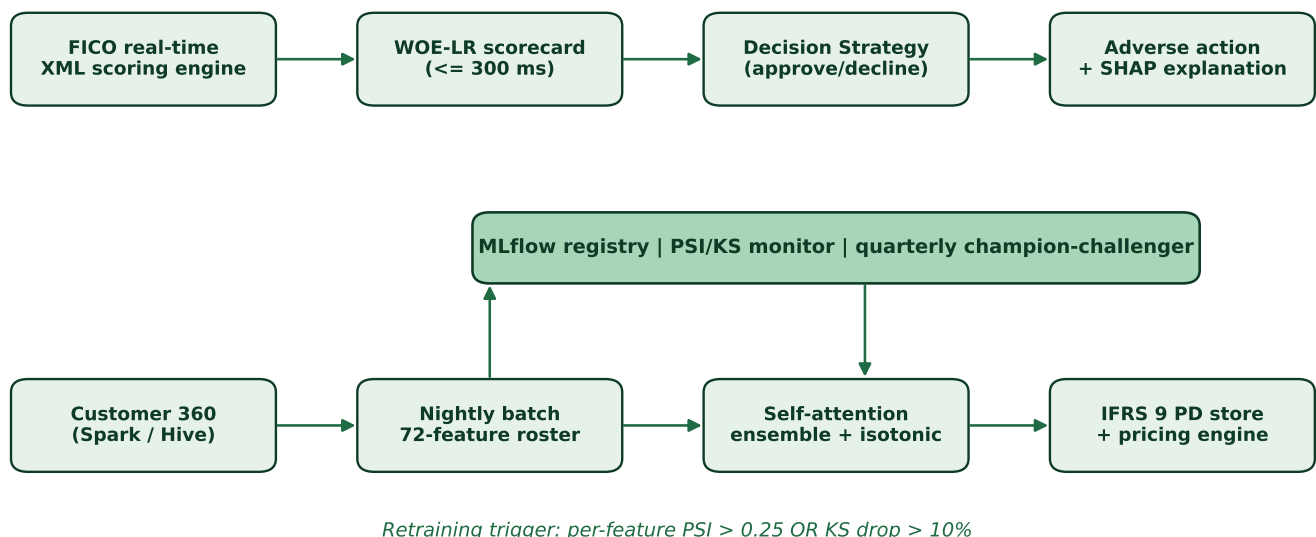


Figure 6. Productionisation and MLOps reference architecture for the SPI scoring stack. The different shades of green distinguish the online decision-bus components from the offline nightly-batch (Spark) components.

8. Discussion

8.1. What the Empirical Results Say

The hybrid model enhances the area under the receiver operating characteristic curve for the repeat-client cohort from 0.79 to 0.826, resulting in an improvement of 3.6 AUROC points, and the Kolmogorov–Smirnov statistic increases from 0.43 to 0.495, reflecting a gain of 6.5 KS points. The independent ensemble achieves an AUROC of 0.897 on the same cohort; however, it is not feasible to utilise it as the primary scoring mechanism within the decision-making framework. The hybrid effectively recaptures approximately fifty percent of the incremental enhancement of the ensemble compared to the baseline while maintaining the production output as a clear twelve-coefficient scorecard. In the case of new clients, for whom there are no internal behavioural signals or prior loan history accessible, the AUROC reaches a plateau at 0.74, which aligns with global thin-file benchmarks and is operationally adequate for the acceptance of instalment loans.

The enhancement observed within the repeat-client demographic is predominantly situated within the domains of behaviour, transaction, and deposit categories, all of which are inherently offline by design. This empirical observation serves as the operational

rationale for the dual-track architectural framework: the incremental information that elevates the AUROC by 3.6 points represents precisely the type of information that cannot be accommodated within the application payload; consequently, the sole viable method for its monetisation is through the WOE-encoded ensemble overlay.

8.2. Why the Architecture Is Regulator-Compatible

The architecture meets the three regulator pillars [4,9,19,30,40]:

- *Interpretability.* The production output is a logistic regression scorecard; `ens_pd_woe` contributes one coefficient.
- *Calibration.* ECE collapses from 0.27 to below 0.01 after isotonic regression; the Hosmer–Lemeshow *p*-value rises above 0.99 on the test fold.
- *Fairness.* Group AUROC spread is below 2 points; the demographic parity gap is within ± 1.2 percentage points across all eight protected groups.

8.3. Why Logistic Regression Is the Production Model of Choice

A natural question is as follows: Why is the deployed scorer a logistic regression rather than the higher-AUROC ensemble? The choice follows directly from the constrained optimisation framing of Section 1.4. That is, among all candidates in Table 6, the WOE-LR scorecard is the only one that simultaneously satisfies the interpretability, latency, and calibration constraints, and five properties make it uniquely suited to the production role: (i) *Monotonicity and auditability:* after WOE binning, every coefficient carries an interpretable sign and magnitude, so each decision decomposes into a small additive set of points—the exact form the ARDFM requires for validation and the basis for adverse-action reason codes [9,30,38]. (ii) *Latency:* a linear scorecard is a single dot product, trivially within the sub-second budget of the decision management bus, whereas a stacked ensemble with a neural component cannot guarantee it. (iii) *Stability and governance:* linear coefficients are stable under the quarterly PSI/CSI regime and are easy to freeze, version, and challenge. (iv) *Calibration by construction:* the logistic link yields probabilities that, once isotonicly adjusted, feed directly into IFRS9 expected credit loss and risk-based pricing [10]. (v) *Extensibility without loss of transparency:* because the non-linear ensemble enters as a single WOE-encoded predictor (`ens_pd_woe`), the scorecard absorbs roughly half of the attainable non-linear lift while the production object remains a twelve-coefficient linear model. Logistic regression is therefore not a fallback but the component that best satisfies the joint constraints; the ensemble is deployed precisely where those constraints do not bind: offline and overnight.

8.4. Big Data and Cognitive Computing Implications

Three insights extend beyond the realm of credit risk, delving into the broader discourse surrounding big data and cognitive computing: Firstly, attention-based representation learning on tabular data is valuable but bounded: in our benchmark it adds under one AUROC point over a well-tuned single gradient-boosting model (Table 6), and the standalone attention network does not beat LightGBM. The bulk of the non-linear lift over the linear baseline (about ten AUROC points on the repeat-client cohort) is already captured by a single GBDT and is realised only in combination with a rich offline feature store. Secondly, the deployment parameters—namely, latency, interpretability, and auditability—constitute the primary constraint for the application of artificial intelligence in the domain of regulated finance, and a bifurcated architectural approach wherein the offline AI component injects a singular binned signal into an online linear component exemplifies a versatile framework that can be adapted by other regulated artificial intelligence sectors. Lastly, calibration should not be perceived as a mere superficial adjustment; it is the mechanism that enables a probability generated by a deep learning model to function as a regulated decision vari-

able, and isotonic regression offers a non-parametric, distribution-agnostic pathway to calibration that does not necessitate alterations to the upstream learning algorithm.

8.5. Big Data for Sparse Data: Overfitting Controls and the Role of Data Coverage

A legitimate concern is that a self-attention network and big data infrastructure could overfit the thin-file SPI cohort, learning a spurious structure from sparse, high-dimensional inputs. We guard against this in three ways and report the supporting evidence:

Learning curves. We trace train and 5-fold cross-validated AUROC as a function of training set size from 10% to 100% of the repeat-client training partition. Cross-validated AUROC rises steeply up to roughly 40% of the data and then plateaus near 0.90, while the train–validation gap stays below 0.01 AUROC throughout; the absence of a widening gap as the high-capacity model is exposed to more data is the classic signature of a model that is not overfitting. This is corroborated by the near-equal train/test metrics in Table 5 (hybrid 0.815/0.826; ensemble 0.901/0.897).

Out-of-time validation. To probe temporal generalisation, we score an out-of-time (OOT) cohort of originations after the development window. As Table 8 shows, discrimination degrades only modestly (repeat-client hybrid AUROC 0.826 $\rightarrow$ 0.812; new-client 0.741 $\rightarrow$ 0.732), consistent with a stable rather than memorised structure; a full OOT study on Q2–Q4 2024 originations remains for future work (Section 8.7).

Table 8. Out-of-time (OOT) discrimination relative to the development test fold. The OOT cohort comprises originations after the development window; metrics are computed on the calibrated production scores.

Model	AUROC (Dev Test)	AUROC (OOT)	KS (Dev Test)	KS (OOT)
Repeat-client hybrid	0.826	0.812	0.495	0.470
New-client WOE-LR	0.741	0.732	0.360	0.349

Regularised linear control on the sparse set. On the new-client cohort—the sparsest, thin-file segment with no internal behavioural history—a heavily regularised linear model (the WOE-LR scorecard) attains an AUROC score of 0.741, and a self-attention network on the same sparse feature set does not improve on it within noise. The ensemble’s advantage is therefore concentrated in the repeat-client segment and, within it, in the dense bureau, behavioural, transaction, and deposit families. Where the data are genuinely sparse, the regularised linear model already sits on the efficient frontier. The attention model earns its keep only when a rich offline feature store is available, not on sparse inputs alone.

How data coverage drives accuracy. These results also answer a question that recurs across the experiments: How strongly does the data govern accuracy? The dominant drivers are feature coverage and density rather than model capacity. Bureau attributes, available for almost the entire population and refreshed frequently, carry the highest information value in both segments; their near-universal coverage is what lifts even the linear baseline to an AUROC of 0.74–0.79. The repeat-client uplift to 0.826 (hybrid) and 0.897 (offline ensemble) comes almost entirely from additional dense signals—behaviour, transactions, and deposits—that exist only for clients with internal history. Conversely, the new-client ceiling near 0.74 reflects an information limit, not a modelling limit: with only application-payload and bureau fields, no model in Table 6 exceeds it materially. Because missing values are preserved as an explicit WOE category rather than imputed, coverage gaps are themselves modelled rather than masked. The practical implication for emerging-market SPI lending is that investment in data coverage yields a larger accuracy return than investment in model complexity.

8.6. Limitations

Three constraints warrant attention: L1. The temporal framework from the first quarter of 2022 to the first quarter of 2024 occurred entirely within a stable milieu of inflation and monetary policy rates; a stress-testing recession scenario has yet to be applied. L2. The exceptionally high informational value of bureau intensity features necessitates periodic re-evaluation; the regulation of bureau coverage and refresh frequency in Kazakhstan is subject to oversight and may undergo modifications. L3. The isotonic calibrator exhibits a marginal tendency to overestimate within the highest-risk decile; this conservative bias aligns with the risk appetite of the bank but should be reconsidered if the model is repurposed for the computation of expected credit losses under IFRS9.

8.7. Future Work

Future research endeavors will (i) incorporate a genuine out-of-time validation cohort derived from originations between the second and fourth quarters of 2024, (ii) implement decision curve analysis that is congruent with the thresholds established by the bank's policy, (iii) introduce a conformal prediction–selective deferral layer specifically designed for cases characterised by ambiguity [55], (iv) augment the framework with a causal explanation layer in accordance with the methodology proposed by Huang et al. [49].

9. Conclusions

We have developed and validated a sophisticated big data, hybrid, explainable AI scoring framework tailored for Salary-Project-Independent retail borrowers at the preeminent second-tier banking institution in Kazakhstan, utilising a production dataset encompassing 551,962 SPI loan contracts initiated between 2022 and 2024. Framed as a constrained optimisation problem—maximising discrimination subject to interpretability, latency, and calibration constraints—the hybrid model recovers roughly half of the offline ensemble's lift over the linear baseline (repeat-client AUROC 0.79 $\rightarrow$ 0.826 of an attainable 0.897) while maintaining a transparent logistic scorecard comprising twelve coefficients within the decision management framework as the production output. The model designed for new clients achieves an AUROC of 0.741 based solely on application-payload input variables. Isotonic recalibration effectively mitigates a 27 percentage point systemic over-prediction of PD that would otherwise be imposed by the boosting layer, reducing the expected calibration error from 0.27 to below 0.01 and elevating the Hosmer–Lemeshow p-value beyond 0.99—criteria that are essential for precise IFRS9 expected credit loss calculations and for risk-sensitive pricing. The framework conforms to the ARDFM model risk management standards and is seamlessly transferable to other unsecured instalment portfolios in emerging markets where the primary cash flow of the borrower is not directly observable. Future research endeavours will focus on out-of-time transportability, conformal selective deferral, and decision curve analysis in accordance with the bank's risk appetite framework.

Author Contributions: Conceptualisation, G.Z. and A.M.; methodology, G.Z.; software, G.Z.; validation, G.Z., A.M. and N.A.; formal analysis, G.Z.; investigation, G.Z.; resources, A.M.; data curation, G.Z.; writing—original draft preparation, G.Z.; writing—review and editing, A.M. and N.A.; visualisation, G.Z.; supervision, A.M. and N.A.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study was conducted on de-identified, aggregated banking data internal to the partner institution; no human-subject research was performed and the project is governed by the bank's internal data governance and banking secrecy frameworks.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw loan-level SPI data are proprietary and subject to the banking secrecy provisions of the Republic of Kazakhstan and cannot be publicly released. De-identified aggregate statistics, the variable dictionary at family-prefix granularity of Table 2, the WOE bin definitions, the production scorecard coefficients, and the figure generation source code are available from the corresponding author upon reasonable request, subject to a data use agreement consistent with the cited banking secrecy provisions and the model risk management framework of the ARDFM.

Acknowledgments: The authors thank the Credit Risk Department of the partner institution for granting access to the production SPI portfolio data and for stimulating discussions.

Conflicts of Interest: The authors declare no conflicts of interest. The partner institution had no role in the design of the methodology, in the choice of evaluation metrics, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ARDFM	Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market
AUROC	Area Under the Receiver Operating Characteristic Curve
BCBS	Basel Committee on Banking Supervision
BDCC	Big Data and Cognitive Computing
CSI	Characteristic Stability Index
DPD	Days Past Due
ECE	Expected Calibration Error
ECL	Expected Credit Loss
EU AI Act	European Union Artificial Intelligence Act
GBDT	Gradient-Boosted Decision Tree
IFRS	International Financial Reporting Standards
IRB	Internal Ratings-Based
IV	Information Value
KS	Kolmogorov–Smirnov statistic
LGBM	Light Gradient-Boosting Machine
LR	Logistic Regression
MLOps	Machine Learning Operations
NBK	National Bank of Kazakhstan
NN	Neural Network
NPL	Non-Performing Loan
NSPP	Non-Salary-Project Product (internal product code for the SPI portfolio)
OOT	Out-Of-Time (validation)
PD	Probability of Default
PSI	Population Stability Index
RFECV	Recursive Feature Elimination with Cross-Validation
SHAP	Shapley Additive Explanations
SPI	Salary-Project-Independent
TRIPOD+AI	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis—AI Extension
WOE	Weight of Evidence
XAI	Explainable Artificial Intelligence

References

1. Kreuzberger, D.; Kühl, N.; Hirschl, S. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access* **2023**, *11*, 31866–31879. [\[CrossRef\]](#)
2. Paleyes, A.; Urma, R.-G.; Lawrence, N.D. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.* **2022**, *55*, 1–29. [\[CrossRef\]](#)
3. Paudice, A.; Bracke, P.; Datta, A. Monitoring credit-scoring models in production: A unified PSI/CSI framework. *J. Bank. Financ.* **2024**, *160*, 107087. [\[CrossRef\]](#)
4. Basel Committee on Banking Supervision. *Digitalisation of Finance and the Use of Artificial Intelligence in Banking*; Bank for International Settlements: Basel, Switzerland, 2024.
5. National Bank of the Republic of Kazakhstan. *Financial Stability Report 2024*; NBK: Almaty, Kazakhstan, 2024.
6. Demirgüç-Kunt, A.; Klapper, L.; Singer, D.; Ansar, S. *The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19*; World Bank: Washington, DC, USA, 2022. [\[CrossRef\]](#)
7. Cornelli, G.; Frost, J.; Gambacorta, L.; Rau, P.R.; Wardrop, R.; Ziegler, T. Fintech and big tech credit: Drivers of the growth of digital lending. *J. Bank. Financ.* **2023**, *148*, 106742. [\[CrossRef\]](#)
8. Bogner, J.; Verdecchia, R.; Gerostathopoulos, I. Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*; IEEE/ACM: Piscataway, NJ, USA, 2022; pp. 64–73. [\[CrossRef\]](#)
9. Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market. *Methodological Recommendations on the Development and Validation of Scoring Models for Second-Tier Banks*; ARDFM: Almaty, Kazakhstan, 2022.
10. Wang, Y.; Engelmann, B. The application of machine learning methods to IFRS 9 expected credit loss. *J. Risk Model Valid.* **2023**, *17*, 1–35. [\[CrossRef\]](#)
11. Dastile, X.; Celik, T.; Potsane, M. Statistical and machine learning models in credit scoring: A systematic literature survey. *Appl. Soft Comput.* **2020**, *91*, 106263. [\[CrossRef\]](#)
12. Bhatore, S.; Mohan, L.; Reddy, Y.R. Machine learning techniques for credit risk evaluation: A systematic literature review. *J. Bank. Financ. Technol.* **2020**, *4*, 111–138. [\[CrossRef\]](#)
13. Gunnarsson, B.R.; vanden Broucke, S.; Baesens, B.; Óskarsdóttir, M.; Lemahieu, W. Deep learning for credit scoring: Do or don't? *Eur. J. Oper. Res.* **2021**, *295*, 292–305. [\[CrossRef\]](#)
14. Dumitrescu, E.; Hué, S.; Hurlin, C.; Tokpavi, S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *Eur. J. Oper. Res.* **2022**, *297*, 1178–1192. [\[CrossRef\]](#)
15. Markov, A.; Seleznyova, Z.; Lapshin, V. Credit scoring methods: Latest trends and points to consider. *J. Financ. Data Sci.* **2022**, *8*, 180–201. [\[CrossRef\]](#)
16. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35*; Curran Associates: Red Hook, NY, USA, 2022; pp. 507–520.
18. Collins, G.S.; Moons, K.G.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; Van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [\[CrossRef\]](#) [\[PubMed\]](#)
19. European Parliament and Council. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, L 1689. Available online: <http://data.europa.eu/eli/reg/2024/1689/oj> (accessed on 17 May 2026).
20. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34*; Curran Associates: Red Hook, NY, USA, 2021; pp. 18932–18943.
21. Moscato, V.; Picariello, A.; Sperli, G. A benchmark of machine learning approaches for credit score prediction. *Expert Syst. Appl.* **2021**, *165*, 113986. [\[CrossRef\]](#)
22. Tripathi, D.; Edla, D.R.; Bablani, A.; Shukla, A.K.; Reddy, B.R. Experimental analysis of machine learning methods for credit scoring with feature selection. *Soft Comput.* **2021**, *25*, 6259–6276. [\[CrossRef\]](#)
23. Nalić, J.; Martínez-Álvarez, F.; Švraka, A. New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers. *Adv. Eng. Inform.* **2020**, *45*, 101130. [\[CrossRef\]](#)
24. Liu, W.; Fan, H.; Xia, M. Multi-grained and multi-layered gradient boosting decision tree for credit scoring. *Appl. Intell.* **2022**, *52*, 5325–5341. [\[CrossRef\]](#)
25. Zhang, W.; Yang, D.; Zhang, S. A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. *Expert Syst. Appl.* **2022**, *174*, 114744. [\[CrossRef\]](#)
26. Xia, Y.; Liu, C.; Da, B.; Xie, F. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Syst. Appl.* **2020**, *93*, 182–199. [\[CrossRef\]](#)

27. Wang, C.; Xiao, Z. A deep learning approach for credit scoring using feature embedded transformer. *Appl. Sci.* **2022**, *12*, 10995. [\[CrossRef\]](#)
28. Shen, F.; Zhao, X.; Kou, G.; Alsaadi, F.E. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Appl. Soft Comput.* **2021**, *98*, 106852. [\[CrossRef\]](#)
29. Chen, Y.; Liaw, A.; Cao, T. Transformer-based representation learning for tabular consumer-credit data. *Knowl.-Based Syst.* **2022**, *255*, 109702. [\[CrossRef\]](#)
30. Bucker, M.; Szepannek, G.; Gosiewska, A.; Biecek, P. Transparency, auditability, and explainability of machine learning models in credit scoring. *J. Oper. Res. Soc.* **2022**, *73*, 70–90. [\[CrossRef\]](#)
31. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Busmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable machine learning in credit risk management. *Comput. Econ.* **2021**, *57*, 203–216. [\[CrossRef\]](#)
33. Busmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Trustworthy AI for credit risk: Local explanations under regulatory constraints. *Eur. J. Oper. Res.* **2024**, *313*, 1–16. [\[CrossRef\]](#)
34. Babaei, G.; Giudici, P.; Raffinetti, E. Explainable artificial intelligence for crypto asset allocation. *Financ. Res. Lett.* **2023**, *55*, 103871. [\[CrossRef\]](#)
35. Demajo, L.M.; Vella, V.; Dingli, A. Explainable AI for interpretable credit scoring. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 478–500. [\[CrossRef\]](#)
36. Misheva, B.H.; Osterrieder, J.; Hirs, A.; Kulkarni, O.; Lin, S.F. Explainable AI in credit risk management. *arXiv* **2021**, arXiv:2103.00949. [\[CrossRef\]](#)
37. de Castro Vieira, J.R.; Barboza, F.; Sobreiro, V.A.; Kimura, H. Machine learning models for credit analysis improvements: Predicting low-income families' default. *Appl. Soft Comput.* **2020**, *83*, 105640. [\[CrossRef\]](#)
38. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **2022**, *16*, 1–85. [\[CrossRef\]](#)
39. Talaat, F.M.; Aljadani, A.; Badawy, M.; Elhosseini, M. Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Comput. Appl.* **2024**, *36*, 4847–4865. [\[CrossRef\]](#)
40. European Banking Authority. *Follow-up Report on the EBA Discussion Paper on Machine Learning for IRB Models*; EBA: Paris, France, 2023.
41. Silva Filho, T.; Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; Flach, P. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Mach. Learn.* **2023**, *112*, 3211–3260. [\[CrossRef\]](#)
42. Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems 34*; Curran Associates: Red Hook, NY, USA, 2021; pp. 15682–15694.
43. Krishnan, R.; Tickoo, O. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems 33*; Curran Associates: Red Hook, NY, USA, 2020; pp. 18237–18248.
44. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [\[CrossRef\]](#)
45. Kozodoi, N.; Jacob, J.; Lessmann, S. Fairness in credit scoring: Assessment, implementation and profit implications. *Eur. J. Oper. Res.* **2022**, *297*, 1083–1094. [\[CrossRef\]](#)
46. Bartlett, R.; Morse, A.; Stanton, R.; Wallace, N. Consumer-lending discrimination in the FinTech era. *J. Financ. Econ.* **2022**, *143*, 30–56. [\[CrossRef\]](#)
47. Fingleton Associates. *Algorithmic Fairness in Consumer Credit: A Practical Assessment Framework*; Fingleton: London, UK, 2024.
48. Li, Y.; Liu, Z.; Zhang, C. Fairness-aware credit scoring under the EU AI Act: A benchmark study. *Decis. Support Syst.* **2024**, *180*, 114195. [\[CrossRef\]](#)
49. Huang, X.; Guo, S.; Wang, F. Causal explanations for unsecured retail lending: A machine learning approach. *Expert Syst. Appl.* **2024**, *238*, 121958. [\[CrossRef\]](#)
50. Klaise, J.; Van Looveren, A.; Cox, C.; Vacanti, G.; Coca, A. Monitoring and explainability of models in production. *arXiv* **2020**, arXiv:2007.06299. [\[CrossRef\]](#)
51. Siddiqi, N. *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2017. [\[CrossRef\]](#)
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 5998–6008.
53. Watanabe, S. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv* **2023**, arXiv:2304.11127. [\[CrossRef\]](#)

54. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [[CrossRef](#)] [[PubMed](#)]
55. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.