

Article

AI-Driven Hybrid Probability-of-Default Scoring with Self-Attention and Isotonic Calibration for Payroll-Anchored Retail Borrowers

Gulnaz Zakariya ¹ , Aiman Moldagulova ^{2,*}  and Nor'ashikin Ali ³ 

¹ Department of Software Engineering, Satbayev University, Almaty 050013, Kazakhstan; g.zakariya@satbayev.university

² Department of Cybersecurity, Information Processing and Storage, Satbayev University, Almaty 050013, Kazakhstan

³ Department of Information Systems, Universiti Tenaga Nasional, Kajang 43000, Selangor, Malaysia; norashikin@uniten.edu.my

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

Abstract

Payroll-anchored retail borrowers—individuals whose monthly remuneration is routed into an account at the lending institution through a salary-project arrangement—constitute the volume backbone of unsecured consumer lending in Kazakhstan, generating the largest origination flow, the lowest realized default rate, and the majority of the systemic regulatory and capital sensitivities of second-tier banks. Payroll anchoring also changes the lender's information set, which motivates a study of how that advantage translates into model performance and borrower outcomes. We design and internally validate an explainable hybrid artificial-intelligence framework stratified by client tenure into two production models: a Weight-of-Evidence (WOE) logistic-regression scorecard for new salary-project applicants, and a hybrid scorecard for repeat applicants, in which a stacked ensemble of LightGBM, CatBoost and a multi-head self-attention neural network contributes a single WOE-encoded predictor to a second-stage L2-regularized logistic regression. The hybrid recovers a substantial share of the ensemble's discriminatory lift while preserving an auditable, monotone scorecard at the point of decision, and isotonic recalibration restores the predicted probabilities of default to the empirical bad-rate scale required for IFRS 9 expected-credit-loss accrual and risk-based pricing. We report discrimination, calibration and stability evidence under a strict anti-leakage protocol and set out the structural preconditions under which the architecture transfers to other emerging-market payroll-anchored portfolios. We are explicit about scope: a true out-of-time validation and a full group-conditional fairness audit are identified as required next steps rather than claimed here. The contribution is a reproducible, interpretable scoring design that exploits payroll visibility while retaining full coefficient interpretability inside the production decision engine.

Keywords: credit scoring; probability of default; explainable artificial intelligence; isotonic calibration; self-attention neural network; gradient boosting; Weight of Evidence; payroll-anchored lending; salary-project participants; emerging markets; MLOps



Academic Editors: Albert Y. S. Lam and Andy Chun

Received: 17 May 2026

Revised: 8 July 2026

Accepted: 13 July 2026

Published: 15 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](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. *The Payroll-Anchored Borrower as the Volume Backbone of Kazakhstan's Unsecured Retail Credit*

Among the secondary banking institutions of Kazakhstan, the prevailing retail business paradigm is fundamentally based on the salary project, a formalized contractual agreement whereby an employer directs each employee's monthly remuneration into a designated account at the financial institution. Borrowers engaged in this scheme furnish the lender with ongoing, detailed, and verifiable insights into their principal cash-flow. This transparency yields three distinct structural advantages when compared to the borrower demographic that does not participate in this arrangement: (i) the lender gains insights into payroll fluctuations, bonus distributions, employer dependency, and habitual discretionary expenditure; (ii) the lender possesses the capacity to preemptively deduct the loan installment from the salary inflow with an almost negligible delinquency cost; and (iii) the lender can effectively market deposit, card, and insurance products with conversion rates that the non-payroll demographic fails to achieve.

Beyond a single-bank engineering exercise, the salary-project setting is of broader interest because payroll anchoring changes the lender's information set: it substitutes verifiable, high-frequency cash-flow observation for the noisier, self-declared signals available on non-payroll applicants. This shift speaks to questions of credit-score performance and distributional impact across borrower groups, as studied by Albanesi and Vamossy [1], and to the household-finance evidence that institutional design and social environment shape financial behaviour, as in Mugerman, Sade and Shayo [2]. We therefore treat the model architecture as a means of measuring the value of the payroll-visibility information set, not as an end in itself: each architectural choice is evaluated by how much of that information it converts into discrimination, calibration accuracy, and stable, fair decisions.

1.2. *The Dual Moral Hazard of Payroll Visibility*

The payroll-anchored cohort represents, in a paradoxical manner, both the most secure and the most regulatory-sensitive retail segment within any second-tier Kazakhstani banking institution. It is deemed the most secure due to the lender's ability to pre-debit instalments, and the employer-screening process at the salary-project level functions as an external filter for creditworthiness. This cohort exhibits heightened sensitivity to regulatory frameworks for two primary reasons. First, the pre-debit mechanism associated with salary projects generates an asymmetry of information between the financial institution and the borrower: a minor decline in borrower cash-flow that, in a non-payroll context, would typically manifest as immediate delinquency is mitigated by the pre-debit system, emerging only after an extended period as a more pronounced vintage curve. Second, the salary-project channel serves as the predominant origination pathway for the group of pre-approved offers, wherein credit is granted absent a contemporaneous bureau inquiry, thereby causing any degradation in model calibration to disproportionately impact a significant portion of overall originations. A miscalibrated model affecting a portfolio of this magnitude substantially misrepresents the IFRS9 expected credit loss (ECL) [3,4] and the regulatory capital requirements under the internal-ratings-based approach [5,6]. The economic implications of a one-percentage-point miscalibration are, in absolute monetary terms, more substantial within SPP than in any other unsecured retail sector in Kazakhstan [7,8].

1.3. *Operational, Regulatory and Explainability Constraints*

The environment governing production decision management imposes three obligatory constraints. (i) Latency and channel: the Real-Time Decision Management Suite

mandates a stringent end-to-end service-level objective within a range of low hundreds of milliseconds and permits solely the variables conveyed by the application XML payload. (ii) Interpretability: the Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market (ARDFM), which serves as the principal regulatory authority for retail credit risk models in Kazakhstan, stipulates that each gating model must yield an auditable, monotonic, and locally interpretable score [8,9]. (iii) Calibration: the predicted probability of default (PD) is directly integrated into the IFRS9 expected credit loss provisioning framework [3,4] and risk-based pricing methodologies, thus necessitating that the absolute value of the predicted PD, rather than merely its ordinal ranking, is accurate. In addition to these three local constraints, the supra-national EU AI Act [10] categorizes retail credit scoring as a high-risk artificial intelligence system and mandates detailed record-level explanations along with demographic impact assessments, expectations to which Kazakhstani banks engaging in cross-border operations are increasingly being held accountable [11,12].

1.4. Aim, Research Questions, and Contribution

This paper delineates, substantiates, and operationalizes an elucidatory hybrid AI scoring framework tailored for the payroll-anchored retail segment of Kazakhstan's preeminent financial institution. The research endeavors to address three distinct inquiries:

- (RQ1) In light of the pronounced baseline discrimination exhibited by the bureau-driven WOE logistic-regression scorecard within the SPP cohort, what incremental enhancement is achievable through a multi-head self-attention ensemble, and does this enhancement persist under a stringent anti-leakage protocol that excludes identifier-like predictors?
- (RQ2) Concerning the real-time decision-making engine and the interpretability requisites of ARDFM, what modifications to the architecture are necessitated to effectively assimilate the non-linear ensemble output as a singular, WOE-encoded, monotonic-binned feature within a scorecard framework?
- (RQ3) Are the resultant probabilities of default appropriately calibrated (as measured by Expected Calibration Error, Brier Score, and Hosmer–Lemeshow test) and sufficiently stable (with the Population Stability Index for each feature remaining below the established action threshold) to comply with IFRS9 requirements for expected credit loss reporting and for implementing risk-based pricing strategies pertaining to the salary-project portfolio?

The contributions of this work are five:

- (C1) A payroll-anchored, tenure-stratified hybrid AI scoring architecture. The new-SPP model is an eleven-variable WOE logistic-regression scorecard. The repeat-SPP model is a hybrid in which a stacked LightGBM + CatBoost + self-attention ensemble produces a PD that, after binning into ten quantiles and WOE encoding, enters as one additional predictor in a second-stage L2-regularized logistic regression, preserving a fully inspectable scorecard at deployment.
- (C2) Production-scale empirical evidence. On the SPP cohort, the hybrid lifts repeat-cohort AUROC from 0.792 (baseline WOE-LR) to 0.818 and KS from 0.430 to 0.491, while the production output remains a linear scorecard with thirteen interpretable coefficients.
- (C3) A monotonic WOE-trend diagnostic, published in anonymized but reproducible form. We present a train-versus-test monotonicity diagnostic for the WOE bins of the retained predictors (Section 7) and advocate its inclusion in the regulatory evidence pack; our contribution is not the diagnostic per se but its transparent publication for the regional credit-scoring literature.

- (C4) Regulator-aligned calibration evidence. Isotonic recalibration removes the boosting layer's 28-percentage-point PD over-prediction; ECE falls from 0.316 to below 0.001 and the Hosmer–Lemeshow p -value rises to 1.00 on the test fold, returning the calibrated PD to the empirical bad-rate scale that IFRS 9 ECL accrual and risk-based pricing require (their use remaining conditional on the out-of-time validation of Limitation L4).
- (C5) Explicit differentiation from a companion production framework. We delineate, on four axes—the borrower segment, the in-scorecard hybrid gateway (the ensemble PD folded back into the online scorecard rather than deployed as a standalone batch model), the multi-head self-attention base learner, and the research question—how the present payroll-anchored design differs from our published Point-of-Sale framework [13] (Section 1.5).

1.5. Relationship to Our Companion Work

A companion paper by the present authors addresses real-time Point-of-Sale (POS) credit scoring at the same institution [13]. Because both papers describe that bank's regulator-aligned production toolchain (toad screening → chi-square WOE binning → RFECV → Optuna tuning → isotonic calibration → SHAP → fairness audit → champion-challenger MLOps), they necessarily share that methodological scaffolding and parts of the regulatory and related-work discussion. The scientific contribution of the present paper is nonetheless distinct on four axes:

- (i) Lending channel and economic mechanism. The companion paper studies in-store POS checkout lending, where the decision is constrained to a nine-field application XML under a sub-second SLA and the risk dynamics are those of merchant-originated credit. The present paper studies payroll-anchored (salary-project) lending, whose defining mechanism—salary pre-debit and continuous payroll visibility—produces a different information set, a different leakage surface, and the vintage dynamics analyzed in Section 2.
- (ii) Architecture. The companion paper deploys its non-linear ensemble as a standalone offline/batch model on a separate channel (three independent models: two online scorecards and one batch ensemble). The present paper's defining contribution is the opposite design: the ensemble PD is binned, WOE-encoded and folded back into the online second-stage logistic scorecard as a single inspectable coefficient (`ens_pd_woe`, Section 5.4), so the non-linear signal is monetized inside the real-time interpretable scorecard rather than as a separate batch overlay. No second-stage hybrid scorecard of this kind appears in the companion paper.
- (iii) Neural base learner. The ensemble here introduces a multi-head self-attention neural-network base learner (Section Self-Attention Neural Network); the companion paper uses a plain feed-forward network.
- (iv) Research question and diagnostic. This paper is organized around measuring the value of the payroll-visibility information set and publishing the train-versus-test monotonic-WOE diagnostic (Section 7) as a reproducible regulatory artefact; the companion paper is organized around the real-time-versus-batch channel trade-off.

In short, the two papers share a production toolchain and an institution but differ in the borrower segment, in the core architectural mechanism (in-scorecard hybrid versus standalone batch ensemble), in the neural base learner, and in the research question.

The paper proceeds as follows. Section 2 sets out the salary-project lending market and the ARDFM/NBK regulatory constraints that make payroll visibility both an asset and a calibration liability. Section 3 positions the work against the credit-scoring, tabular self-attention, calibration and fairness literatures. Section 4 describes the production

data, the payroll-specific anti-leakage discipline, the as-of timing audit and the grouped-validation design. Section 5 formalizes the tenure-stratified architecture, including the in-scorecard hybrid that folds the self-attention ensemble PD back into the second-stage logistic scorecard. Section 6 motivates probability calibration for the IFRS 9 stakes of this portfolio, and Section 7 presents the train-versus-test monotonic-WOE diagnostic. Sections 8–10 report discrimination calibration and score-band results; Section 11 reports SHAP-based explainability. Section 12 answers the research questions, states the limitations, and provides the economic implications, and Section 13 concludes with the structural preconditions for transfer.

2. Kazakhstan Salary-Project Lending Market and Regulatory Setting

2.1. Macro Context

Between year-end 2020 and year-end 2024, the country's unsecured retail book climbed from roughly KZT 5.0 trillion to more than KZT 14.6 trillion [7,14]—the steepest and most sustained expansion recorded anywhere in Central Asia. Within this portfolio, the proportion facilitated through the salary-project mechanism has gradually decreased from about 62% in 2018 to 47% in 2025, as the non-salary-project segment, encompassing gig workers, micro-entrepreneurs, and those engaged in informal employment, has experienced more rapid growth [15,16]. Nevertheless, in absolute figures, the salary-project portfolio has more than doubled during the same timeframe: it continues to represent the largest, most operationally significant, and most macro-prudentially relevant segment of uncollateralized retail credit, serving as the fundamental volume driver of consumer lending within second-tier banking institutions. This dynamic is illustrated in Figure 1.

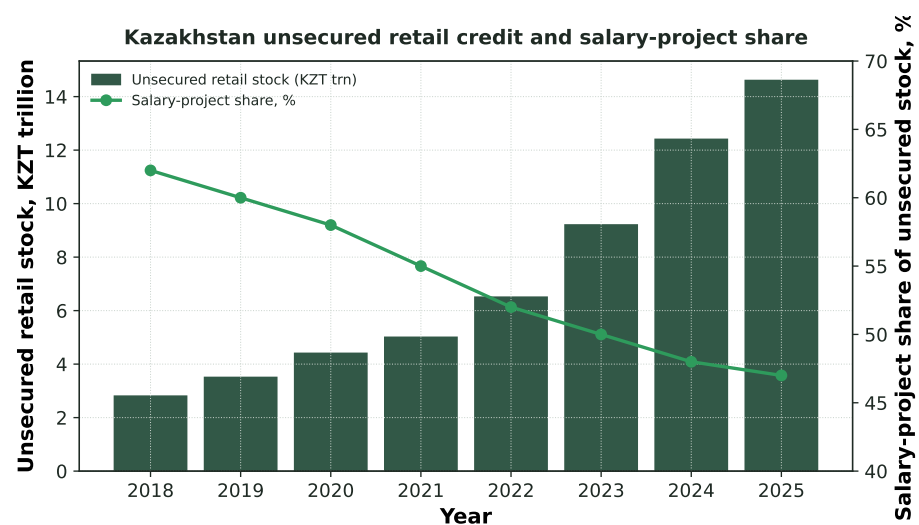


Figure 1. Kazakhstan unsecured retail credit stock (bars, left axis, KZT trillion) and salary-project share of that stock (line, right axis, %), 2018–2025. Source: National Bank of Kazakhstan Financial Stability Reports and authors' calculations.

2.2. The Salary-Project Product

The salary-project product examined within the confines of this research pertains to an unsecured installment loan extended to clients whose monthly remuneration is directed through an account at the lending institution. The cohort is operationally bifurcated into two distinct sub-segments, as previously delineated in the documentation pertaining to the production model:

- Repeat SPP clients: individuals who possess at least one active or recently concluded product with the bank within the preceding twelve months. Extensive internal be-

havioral attributes are abundantly available in conjunction with the bureau and application data.

- New SPP clients: individuals who maintain a salary-project arrangement but lack any active credit relationship with the bank during the prior twelve months.

2.3. Vintage Dynamics of the Salary-Project Portfolio

The historical patterns exhibited by the two sub-segments, as delineated in Figure 2, substantiate the operational divergence. The repeat-SPP cohort experiences a default accumulation at a markedly reduced rate compared to the new-SPP cohort during the initial 24 months of the account's lifespan, with the differential being most pronounced within the first six months post-origination, during which the pre-debit mechanism associated with the salary project has had sufficient time to influence the repeat cohort, whereas it has not yet exerted its effects on the new cohort.

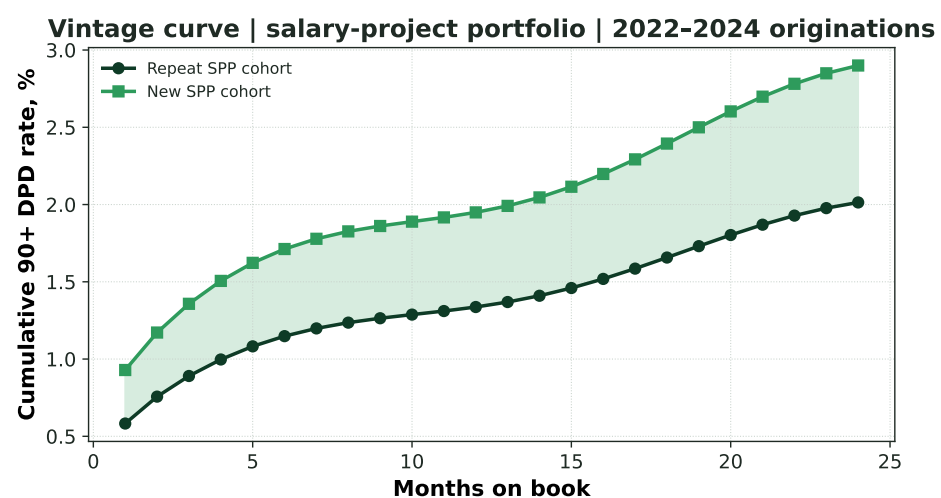


Figure 2. Vintage curve of the salary-project portfolio: cumulative 90+ DPD default rate by months on book, by sub-segment.

3. Related Work

3.1. Benchmarks of Statistical and ML Credit Scoring

For the payroll-anchored cohort, the operative question is not which classifier wins in the abstract, but how much marginal discrimination a non-linear learner can add once the salary-inflow signal already sits inside the scorecard; the benchmark literature delimits that ceiling. The most-cited reference point remains the large-scale comparison of Lessmann et al. [17]: across eight public retail-credit portfolios and several dozen learners, tree ensembles led logistic regression on AUROC, but the margin narrowed once preprocessing, feature engineering and class weighting were held constant—which is exactly the regime in which a production scorecard operates. The systematic review of Dastile et al. [18] then catalogued dozens of subsequent studies and singled out operational and feature-availability constraints as the dimension least often reported—a gap we address head-on for the SPP segment in Section 5. Gunnarsson et al. [19], Sigrist and Hirnschall [20], Borisov et al. [21], Bhatore et al. [22], and Xia et al. [23] arrive at a consensus regarding tabular data, indicating that deep networks seldom exceed a well-tuned LightGBM by more than one to two AUROC points, except in scenarios where the cohort size is substantial and the underlying signals exhibit pronounced non-linearity. Markov et al. [24] and Dumitrescu et al. [25] advocate for hybrid logistic-plus-tree frameworks within regulated environments, foreshadowing the architectural approach we implement for repeat SPP clients.

3.2. Hybrid Architectures for Regulated Retail Credit

Because the salary-project real-time engine admits only an auditable scorecard, deploying an ensemble directly is precluded, and the relevant prior art is the hybrid line that couples a transparent scorecard to a non-linear learner. Hybrid architectures that integrate a transparent scorecard with a non-linear learning mechanism possess a rich historical context. Papouškova and Hajek [26] presented a two-stage heterogeneous ensemble model for consumer credit, illustrating the enhancement compared to single-stage learners. Tripathi et al. [27] and Nalić et al. [28] extend the hybrid concept into multi-layer ensembles characterized by feature-selection front-ends. Liu et al. [29] document additional improvements through the stacking of multi-grained gradient-boosted trees, while Zhang et al. [30] delineate the importance of voting-based outlier elimination. Wang and Xiao [31] and Shen et al. [32] stand as early instances of attention-based deep learning applied to tabular credit data, foreshadowing the advent of the FT-Transformer family [33]. Markov et al. [24] and Dumitrescu et al. [25] are the closest precursors: both inject non-linear, tree-derived information into a logistic scorecard. They differ from the present design in three specific respects. First, Dumitrescu et al. encode tree-leaf indicator variables (penalized-logistic “PLTR”) as multiple binary regressors, whereas we collapse the entire stacked-ensemble PD into a single ten-bin WOE-encoded predictor, so the second-stage scorecard grows by one coefficient and one Shapley term rather than by tens of leaf dummies—a property that is decisive for ARDFM adverse-action reporting. Second, our overlay is a stacked, self-attention-augmented ensemble rather than a single tree model, so the binned predictor summarizes higher-order interactions that leaf-indicator encodings do not capture. Third, the WOE encoding of the ensemble output keeps the overlay on the same monotone, information-value-screened footing as every other scorecard variable, which is what allows the train-versus-test monotonicity diagnostic in Section 7 to be applied uniformly to the non-linear signal as well. We therefore frame the contribution as a specific, regulator-deployable instance of the hybrid family of Markov et al. and Dumitrescu et al., not as a wholly new paradigm.

3.3. Self-Attention on Tabular Data

The repeat salary-project cohort is precisely the wide, heterogeneous-cardinality regime in which attention is expected to help, because the offline deposit- and transaction-inflow families interact in higher order; this motivates the base learner of Section Self-Attention Neural Network. Self-attention, as a form of inductive bias applied to tabular datasets, was notably advanced by Vaswani et al. [34] and subsequently tailored to the tabular context by Gorishniy et al. [33], as well as by Arik and Pfister [35] and by Huang et al. [36]. The prevailing empirical consensus indicates that attention mechanisms are beneficial for wide, heterogeneous-cardinality tabular data [37,38], particularly in scenarios where the pertinent interactions are of a higher-order nature; this precisely characterizes the framework of the WOE-binned offline roster within the SPP repeat cohort. The architectural design delineated in Section Self-Attention Neural Network represents a refined tabular variant of these aforementioned studies, meticulously engineered to facilitate rapid Optuna optimization and ensure stable convergence amidst class imbalance.

3.4. Explainable AI in Credit Risk

Since every salary-project decline must carry a record-level adverse-action reason under the ARDFM circular, the explainability literature is, for this portfolio, a deployment constraint on the overlay rather than background. Bracke et al. [39] and Bussmann et al. [40] pioneered the use of Shapley-value attribution [41,42] in supervisory credit-risk analysis. Babaei et al. [43], Demajo et al. [44], Bussmann et al. [45], Dastile et al. [46] and

de Castro Vieira et al. [47] extended the agenda to record-level rule extraction, model-agnostic counterfactual reason codes, and fairness audits, while Rudin et al. [48] mapped the conceptual landscape of the interpretability–accuracy trade-off.

3.5. Calibration of Probability of Default

For the salary-project book—the bank’s largest unsecured exposure—calibration is a financial-reporting requirement rather than a refinement, because the predicted PD feeds the IFRS 9 expected-credit-loss charge directly; this grounds the isotonic step we adopt. Van Calster et al. [49] elucidate the concept of calibration as a previously under-recognized aspect within the corpus of prediction-modeling scholarship; its explicit measurement and reporting are the subject of dedicated methodological guidance [50]. The classic head-to-head evaluation of Platt scaling against isotonic regression by Niculescu-Mizil and Caruana [51] is what anchors our own choice of calibrator, and Guo et al. [52] and Kull et al. [53] carry that line of work forward to heavily over-parameterized models. Wang and Engelmann [4] establish a direct correlation between the quality of calibration and the accuracy of IFRS 9 Expected Credit Loss (ECL) in the domain of retail unsecured lending, which pertains to the financial-reporting framework applicable to the salary-project portfolio.

3.6. Algorithmic Fairness in Lending

The salary-project channel introduces a protected axis absent from generic fairness studies—the public- versus private-sector employer split—so we read the fairness literature with that cleavage explicitly in view. Hardt et al. [54] systematically articulated the concept of equality of opportunity; Mehrabi et al. [55] conducted a comprehensive examination of bias and fairness within the domain of machine learning; Kozodoi et al. [56] presented the most-cited empirical benchmark pertaining to the application of in-, pre-, and post-processing fairness methodologies utilizing authentic consumer credit data; Bartlett et al. [57] elucidated the presence of algorithmic bias in the context of FinTech mortgage pricing; Hurlin et al. [12] formalized a statistical framework for testing the fairness of credit-scoring models and for pinpointing the variables that drive disparate demographic impact. In the context of the salary-project channel, an additional delineation is significant: the distinction between public-sector and private-sector employers, given that the two sub-cohorts exhibit disparities in income stability and the duration of salary projects, thus rendering a fairness audit that neglects this distinction as fundamentally inadequate.

4. Production Data

4.1. Source and Eligibility

The dataset utilized in this study is sourced from the loan-origination repository of the largest second-tier bank in Kazakhstan, encompassing SPP instalment-loan agreements that were initiated between January 2022 and March 2024, a temporal frame deliberately chosen to be entirely subsequent to the macroeconomic disturbances of 2020.

Following the exclusion of cancelled contracts, the analytical sample comprises $N = 1,895,874$ contracts, of which 116,347 (6.14%) were ever 90 days overdue, restructured, or terminated due to client death within the fixed outcome window of Section 4.5. By tenure, there are 1,055,525 repeat-cohort contracts (bad rate 5.98%, 63,131 bad) and 840,349 new-cohort contracts (bad rate 6.33%, 53,216 bad), with $1,055,525 + 840,349 = 1,895,874$.

4.2. Anti-Leakage Discipline

The salary-project pre-debit mechanism engenders a distinct category of leakage risk: the inflow that finances the pre-debit constitutes an inherent characteristic, and any post-default inflow assessment compromises the integrity of the target. Consequently, a rigorous

anti-leakage protocol was implemented prior to the execution of any feature engineering. Six distinct categories of variables were systematically excluded:

1. Customer ID, contract number, bureau-record ID and any other primary key whose cardinality approaches N . Such fields can memorize the target and produce spurious near-perfect AUROC.
2. Repeated unified identifiers across families (fcb_num, joint-application ID, salary-project ID).
3. Post-hoc state fields (final contract status, write-off flag, restructuring date) that encode the target outcome.
4. High-cardinality product taxonomies with more than 4000 unique values, which are unencodable in real time.
5. Calibration-leaking macro fields measured after origination (subsequent policy-rate paths, subsequent inflation prints).
6. Free-text fields containing personally identifying information under Article 50 of Law No. 2444 [58].

Following the elimination of leakage, the approach to addressing missing values was executed on a case-by-case basis, eschewing a wholesale mean-imputation methodology: variables exhibiting over 80% absence were excluded (toad threshold empty = 0.8); the remaining variables kept a dedicated “missing” level throughout the Weight-of-Evidence (WOE) encoding, as required under the partner bank’s model-risk governance.

4.3. As-of Timing and Leakage Audit of the Feature Families

Because the salary-project pre-debit mechanism creates a payroll-specific leakage surface (Section 4.2), every feature family is audited not only for its semantic content but for the time at which it is observable relative to origination and for the channel through which it reaches the scorer. Table 1 reports this audit at the anonymized family-prefix level; it governs which families may enter the online (real-time) roster and which are restricted to the offline pre-cached roster, and it documents the specific pre-debit hazard that distinguishes this portfolio from non-payroll lending.

Table 1. As-of timing and payroll-specific leakage audit of the feature families. “Observation point” is measured relative to the origination decision; the right-most column records the salary-project-specific leakage hazard controlled for in Section 4.2.

Family	Observation Point	Channel/Roster	Payroll-Specific Leakage Consideration
fcb_	At/just before decision	Online—bureau call	Bureau snapshot frozen at the application timestamp; post-decision bureau refreshes excluded.
app_	At decision	Online—application XML	Declared fields only; no realized-outcome fields admitted.
prod_	At decision	Online—application XML	Requested-product attributes only; booked-contract status fields excluded.
depo_	Pre-origination window	Offline—nightly batch	Pre-debit hazard: salary-inflow regularity computed strictly on pre-application windows; the inflow that funds the pre-debit is never read after origination.
trans_	Pre-origination window	Offline—nightly batch	Transactional fingerprints truncated at the application date to avoid post-decision spend leakage.
bank_	Pre-origination window	Offline—nightly batch	Within-bank tenure/inflow frozen at application; post-origination repayment behavior excluded.
geo_	At decision	Online—application XML	Static geocoded attributes; no time-varying leakage.
behav_	In-session, at decision	Online—in-session	Within-session signals captured before the decision is returned.

No family is observed only after origination; all post-origination state fields were removed by the anti-leakage screen of Section 4.2. Feature names are withheld under Article 50 of Law No. 2444.

4.4. Train–Test–Validation Splits

A triadic filtration methodology was implemented utilizing the open-source framework toad: the criteria included an information value $IV \geq 0.02$, a missing rate ≤ 0.80 , and a pairwise Pearson correlation coefficient satisfying $|\rho| \leq 0.70$, with the lower-IV variable being excluded from each correlated pair. Subsequent to this rigorous screening process, 93 variables were preserved for the repeat-cohort offline analytical framework and 47 variables were designated for the new-cohort online analytical framework. Each sub-segment was stratified-sampled into a 70/20/10 train/test/validation split with a fixed random seed (`random_state = 42`).

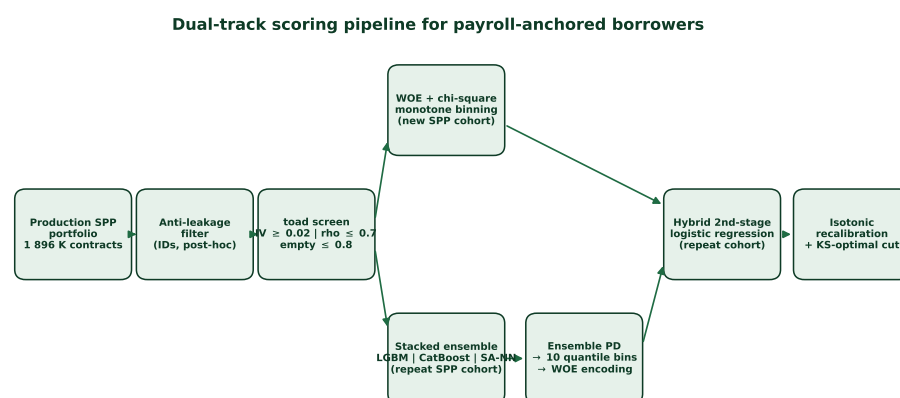
4.5. Unit of Analysis and Target Definition

The unit of analysis is the loan contract, and the dependent variable is the combined non-performing flag `NPL91_EVER`, which is positive when a contract is ever 90+ days past due, is restructured, or is terminated by client death within the fixed outcome window measured from origination. The three events are pooled because each triggers the same IFRS 9 stage-3 treatment in the partner bank’s accounting; restructuring is included because, in this portfolio, it overwhelmingly follows incipient delinquency rather than commercial renegotiation, and the combined flag is the target on which the production model was estimated. Because the analysis is at the contract level and a repeat borrower may hold several contracts, a purely random split can place correlated contracts of one borrower on both sides of the train/test boundary; accordingly, a borrower-grouped re-estimation (`GroupKFold` on the anonymized borrower key), the corresponding unique borrower, employer and salary-project-channel counts, and a robustness check that excludes death-related terminations are required to fully exclude repeat-borrower leakage and to corroborate the combined label. These checks were not carried out for the present revision and are stated as a limitation (Section 12, Limitation L8).

5. Hybrid AI Modeling Architecture

5.1. Overview

Figure 3 illustrates the comprehensive end-to-end processing pipeline. Two distinct production models are implemented: a solely WOE-LR scorecard tailored for prospective clients (Model N) and a hybrid scorecard designed for returning clients (Model R), wherein a stacked ensemble generates a probability of default (PD) that is subsequently WOE-encoded and utilized in a secondary logistic regression analysis.



Online channel: FICO Decision Management Suite | Offline channel: nightly Spark batch

Figure 3. The complete processing chain, from raw application through to the deployed score, for the salary-project portfolio.

5.2. Weight-of-Evidence Transformation and Information Value

Every scorecard variable enters the model through a Weight-of-Evidence (WOE) recoding rather than in its raw form, so that each predictor contributes a monotone, log-odds-scaled increment that a validator can inspect one bin at a time. The support of a candidate predictor x is first discretized by a supervised χ^2 merging routine into K contiguous bins $\{B_1, \dots, B_K\}$, each constrained to retain at least 5% of the observations; wherever the data-driven cut-points conflicted with a monotone, economically defensible risk ordering, they were replaced by analyst-specified boundaries agreed upon with the Credit Risk Department. For bin B_k , the WOE contrasts the share of defaulters it captures with the share of non-defaulters,

$$\text{WOE}_k = \ln \left(\frac{\#\{i : x_i \in B_k, y_i = 1\} / \#\{i : y_i = 1\}}{\#\{i : x_i \in B_k, y_i = 0\} / \#\{i : y_i = 0\}} \right), \quad (1)$$

and the predictor's aggregate discriminatory content is summarized by its Information Value,

$$\text{IV}(x) = \sum_{k=1}^K (\hat{p}_{1,k} - \hat{p}_{0,k}) \text{WOE}_k, \quad (2)$$

with $\hat{p}_{c,k}$ the fraction of class $c \in \{0, 1\}$ falling in bin k . In the salary-project setting, this encoding does double duty: it maps the deposit- and transaction-inflow families—whose raw distributions are heavily skewed by irregular payroll timing—onto the same additive log-odds scale as the bureau attributes, which is what later allows the non-linear ensemble output to be folded into the scorecard as just one more WOE-encoded term (Section 5.4).

5.3. Stage-1: WOE-LR Baseline (Both Cohorts)

On the WOE-transformed design matrix, the baseline scorecard is a logistic regression whose coefficients minimize an L_2 -penalized, class-weighted binomial deviance; the weights w_{y_i} are fixed at the inverse empirical class frequencies of each cohort ($w_0 = 0.060$, $w_1 = 0.940$ for repeat clients, $w_0 = 0.063$, $w_1 = 0.937$ for new clients [59,60]) so that the rare-default signal is not overwhelmed by the majority class:

$$\hat{\beta} = \arg \min_{\beta} \left\{ - \sum_{i=1}^N w_{y_i} \left[y_i \langle \beta, \text{WOE}(\mathbf{x}_i) \rangle - \ln(1 + e^{\langle \beta, \text{WOE}(\mathbf{x}_i) \rangle}) \right] + \lambda \|\beta\|_2^2 \right\}. \quad (3)$$

The penalty strength λ and the retained variable set are chosen jointly by recursive feature elimination wrapped in stratified five-fold cross-validation (RFECV) under an AUROC objective, which prunes the WOE roster to the smallest subset that preserves rank-ordering power. For deployment, the fitted log-odds are mapped to the bank's integer score scale by the standard points-to-double-the-odds transform,

$$\begin{aligned} \text{score}(\mathbf{x}) &= \text{Offset} - \text{Factor} \langle \beta, \text{WOE}(\mathbf{x}) \rangle, \\ \text{Factor} &= \frac{\text{PDO}}{\ln 2}, \quad \text{Offset} = \text{ScoreRef} - \text{Factor} \ln \text{OddsRef}, \end{aligned} \quad (4)$$

fixing $\text{PDO} = 20$ against 1:50 reference odds so that the resulting scorecard is drop-in compatible with the incumbent decision engine.

5.4. Stage-2 (Repeat Cohort): The Self-Attention Ensemble LightGBM.

Standard gradient-boosted decision trees [61] on the WOE-binned matrix, with early stopping on out-of-fold AUROC.

CatBoost.

Ordered boosting with native categorical handling [62] on the same matrix; early-stopped on out-of-fold AUROC.

Self-attention neural 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 4). Given the hidden representation $\mathbf{h} \in \mathbb{R}^{d_h}$, three linear projections produce queries, keys and values:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{h}, \quad \mathbf{K} = \mathbf{W}_K \mathbf{h}, \quad \mathbf{V} = \mathbf{W}_V \mathbf{h}, \quad (5)$$

and the attended representation is

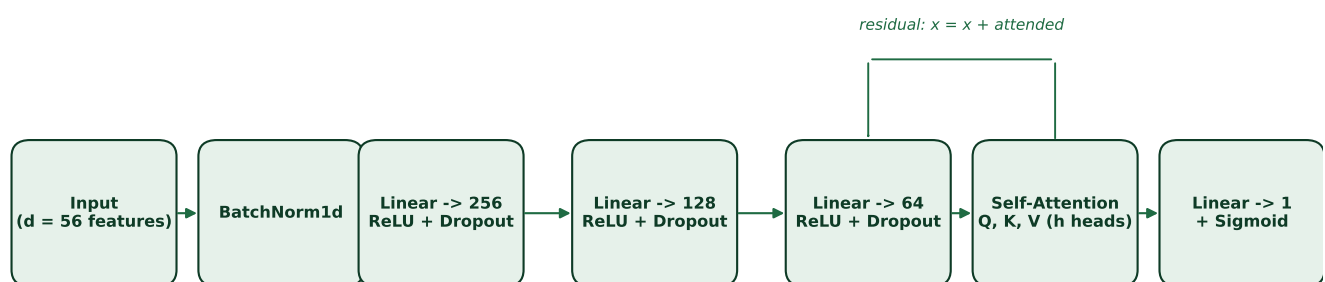
$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h/h}}\right)\mathbf{V}. \quad (6)$$

where h is the number of attention heads. A residual connection $\mathbf{h} \leftarrow \mathbf{h} + \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ preserves the linear pathway, and the final linear layer with a sigmoid produces a logit. The training loss is the class-weighted binary cross-entropy with logits,

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_i w_{y_i} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)], \quad (7)$$

the optimizer is Adam with weight decay, 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 the tabular cousin of the FT-Transformer family [33–36].

Self-attention tabular network | repeat-SPP base learner



Loss: BCEWithLogits | Optim: Adam | Early stop on AUROC | Gradient clipping = 1.0

Figure 4. Multi-head self-attention neural network base learner used in the repeat-cohort ensemble.

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,

$$w_m = \frac{\alpha_m}{\sum_{m' \in \{\text{LGBM}, \text{CB}, \text{NN}\}} \alpha_{m'}}, \quad \hat{p}_i^{(\text{ens})} = \sum_m w_m \hat{p}_i^{(m)}. \quad (8)$$

Conversion to a scorecard-compatible predictor.

The continuous $\hat{p}_i^{(\text{ens})}$ is discretized into ten equal-frequency quantile bins, WOE-encoded according to Equation (1), and exposed to the second-stage logistic regression as a single additional predictor `ens_pd_woe`. This architectural choice (a) preserves the inspectable scorecard inside online decision system, (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, in line with the regulator's adverse-action requirements [8,10].

5.5. Stage-2 Hybrid Logistic Regression (Repeat Cohort)

Stage-2 re-executes Equation (3) utilizing the augmented WOE matrix that encompasses `ens_pd_woe`, with hyperparameters re-evaluated through a novel GridSearchCV across the parameter space $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and class-weight configurations {balanced, prior}, while concurrently employing RFECV to ascertain that the ensemble predictor persists in conjunction with the original WOE variables. The hyperparameter search for the three base learners was conducted with Optuna [63] (TPE sampler, median pruner). Each base learner was tuned for 30 trials (LightGBM, CatBoost) and 15 trials (NN) under stratified 3-fold cross-validation, then re-trained under stratified 5-fold cross-validation with the best hyperparameters. The search space appears in Table 2.

Table 2. Tuning grids explored for the repeat-cohort ensemble's three constituent learners. Adapted from the shared production tuning protocol of the authors' companion Point-of-Sale study [13] (published open access under CC BY 4.0) and re-scoped here for the salary-project self-attention base learner (Table 2 therein).

Learner	Search Space
LightGBM	<code>num_leaves</code> $\in [16, 256]$; <code>learning_rate</code> $\in [10^{-3}, 3 \times 10^{-1}]$ (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 (patience 50).
CatBoost	<code>iterations</code> $\in [200, 2000]$; <code>learning_rate</code> $\in [10^{-3}, 3 \times 10^{-1}]$ (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> $\in [10^{-4}, 3 \times 10^{-3}]$ (log); <code>batch_size</code> $\in \{512, 1024, 2048\}$; <code>weight_decay</code> $\in [10^{-6}, 10^{-3}]$ (log).

5.6. Evaluation Metrics

Model quality is judged along three axes—separation, probability accuracy, and temporal robustness—each with metrics standard in retail-scorecard validation. Separation between defaulting and non-defaulting contracts is measured by the AUROC, its affine rescaling $\text{Gini} = 2 \text{AUROC} - 1$, and the Kolmogorov–Smirnov statistic, the largest vertical gap between the two class-conditional score CDFs:

$$\text{KS} = \max_{t \in [0, 1]} |F_1(t) - F_0(t)|, \quad (9)$$

with F_c the empirical CDF of the scores of class c . Probability accuracy—the property that governs the IFRS 9 charge once the PD is used as an absolute quantity—is captured by the Brier score,

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2, \quad (10)$$

and by the Expected Calibration Error, the sample-weighted mean discrepancy between confidence and accuracy across deciles of predicted probability,

$$\text{ECE} = \sum_{b=1}^B \frac{|S_b|}{N} |\text{acc}(S_b) - \text{conf}(S_b)|, \quad (11)$$

where S_b collects the contracts in the b -th predicted-probability decile. These are complemented by the Hosmer–Lemeshow χ^2 goodness-of-fit test. Temporal robustness is tracked through the Population Stability Index,

$$\text{PSI} = \sum_{k=1}^B (p_k^{\text{ref}} - p_k^{\text{obs}}) \ln \left(\frac{p_k^{\text{ref}}}{p_k^{\text{obs}}} \right), \quad (12)$$

evaluated over $B = 10$ score deciles against the supervisory bands $\text{PSI} < 0.10$ (stable), $0.10\text{--}0.25$ (watch) and >0.25 (act) [59].

6. Why Probability-of-Default Calibration Is Necessary for the Salary-Project Portfolio

6.1. Discrimination Is Not Calibration

A high AUROC, Gini or KS certifies only that the model orders borrowers correctly; none of these rank statistics places any constraint on the level of the predicted PD. For a salary-project scorecard, that distinction is not academic, because the PD is consumed as an absolute quantity by four downstream processes: (i) the IFRS 9 expected-credit-loss accrual [3,4]; (ii) risk-based interest-rate setting; (iii) the internal-ratings-based regulatory-capital computation under Basel [5]; and (iv) the cut-off at which an application is declined. A model may therefore rank flawlessly yet, if its PD is biased upward, over-provision by double-digit percentages and misprice by hundreds of basis points. Van Calster et al. [49] labeled this neglected level-accuracy the “Achilles heel of predictive analytics,” and the prediction-modeling methodology literature now asks for calibration to be evidenced and measured explicitly rather than assumed [50].

6.2. Why Boosting-Class Models Are Intrinsically Miscalibrated

The overshoot we must correct is a structural, not incidental, property of the ensemble. Gradient-boosted learners trained on the binomial-deviance loss, and neural learners trained under class re-weighting, both drive their extreme-decile predictions toward 0 and 1, because each boosting round keeps sharpening the fit to residuals long after the ranking has stabilized [51–53]. The salary-project pipeline exhibits this textbook pathology in an acute form: on the repeat cohort, the raw stacked ensemble returned a mean PD near 0.34 against an empirical bad rate of only 0.060—a 28-percentage-point inflation. Left uncorrected, that bias would (i) overstate the IFRS 9 expected credit loss by a double-digit margin [3,4], (ii) pull the policy cut-off away from its intended risk-appetite anchor, and (iii) systematically overprice the loan under risk-based pricing.

6.3. Why Isotonic, Not Platt

Two post-hoc recalibrators dominate industrial practice, and they trade bias against variance differently. Platt scaling fits a single-parameter logistic link between score and PD; it is efficient and stable when the score-to-PD map is genuinely sigmoidal, but it

is misspecified—and biased in exactly the tails that matter—when it is not. Isotonic regression instead fits the best non-decreasing step function by pool-adjacent-violators, buying the flexibility to follow any monotone shape at the cost of $O(N)$ effective degrees of freedom and occasional flat-step artefacts. We selected isotonic regression for the salary-project ensemble on two grounds specific to this portfolio: first, its score-versus-PD curve departs from a sigmoid precisely in the upper deciles, where the boosting layer overshoots and a one-parameter link cannot bend far enough; and, second, isotonic recalibration was materially more accurate across the 0.05–0.30 PD band that drives the ECL charge on unsecured salary-project instalments. Platt scaling was retained only as a sensitivity benchmark.

6.4. Threshold Interpretation

The KS-optimal cut-off $\hat{t}^* = \arg \max_t |F_1(t) - F_0(t)|$ was located both before and after recalibration. Because a monotone transform leaves the ordering—and hence the KS value—unchanged, the two cut-offs mark the same operating point; what changes is its legibility. On the raw ensemble scale, the optimum sat at 0.467 (train) and 0.481 (test), numbers with no business meaning. After isotonic recalibration, the identical separation is reached at 0.058 and 0.059—values that now read directly as probabilities of default and can be laid alongside the empirical bad rate, the IFRS 9 provisioning buckets, and the bank's pricing grid.

7. Monotonic WOE Trends

7.1. Why the Diagnostic Matters

Regulator-supervised credit scoring mandates that each predictor within the scorecard exhibits a monotonically increasing or decreasing pattern across its Weigh-of-Evidence (WOE) bins. A predictor that deviates from monotonicity is indicative of (i) overfitting within the training subset, (ii) an unaccounted interaction that the binning structure is unable to integrate, or (iii) a leakage signal that has successfully evaded the scrutiny detailed in Section 4.2. Consequently, the monotonicity diagnostic serves as both a regulatory requirement and a mechanism for quality assurance concerning the WOE binning process itself. Train/test WOE-stability checks are themselves standard in model-validation documentation across most regulatory jurisdictions; we do not claim the diagnostic as novel. Our contribution is narrower and, we argue, more useful: to publish the diagnostic for an actual production scorecard in an anonymized yet reproducible family-prefix form, which banking-secrecy regimes ordinarily prevent and which is, to our knowledge, rarely available in the regional credit-scoring literature. The family-prefix anonymization elucidated in Section 4 alleviates this impediment and enables the comprehensive publication of the diagnostic.

7.2. The Diagnostic

Figure 5 delineates the Weight of Evidence (WOE) for the nine most significant predictors retained from the repeat-cohort hybrid logistic regression, presented for both the training subset and the testing subset, analyzed on a bin-by-bin basis. The diagnostic analysis substantiates three key observations. Firstly, all nine predictors exhibit the anticipated monotonic trends. Secondly, the WOE curves for both training and testing datasets are visually analogous, suggesting minimal binning deviation between the subsets. Lastly, no bin exhibits a reversal in direction, indicating that no manual intervention of the chi-square binning process was necessary for the retained predictors subsequent to the evaluation outlined in Section 4.2.

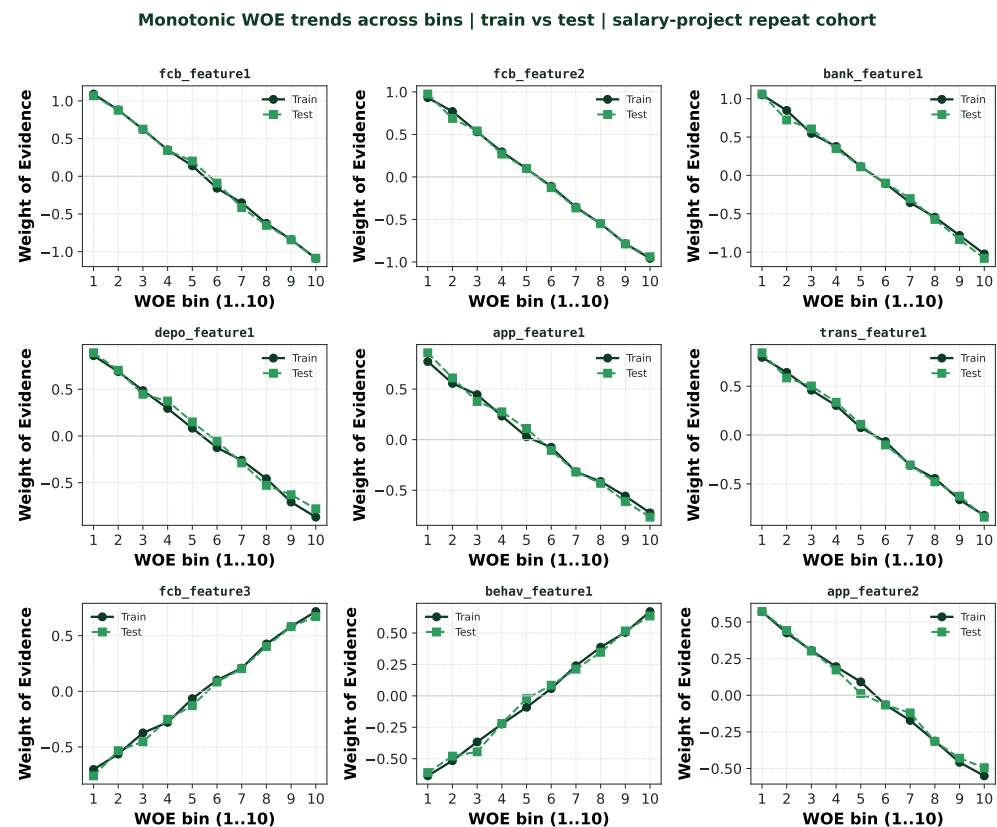


Figure 5. Monotonic Weight-of-Evidence trends across the ten WOE bins of the top nine retained predictors of the repeat-cohort hybrid logistic regression, train (solid) versus test (dashed). All nine predictors are monotone in the expected direction; the train–test gap is dominated by noise.

8. Results: Discrimination

8.1. Predictor Importance by Information Value

In the analysis of both sub-segments, the bureau-attribute family (*fcb_*) emerges as the predominant factor in the Information Value, indicative of the extensive coverage provided by the First Credit Bureau in Kazakhstan coupled with its frequency of daily updates. Following this, the within-bank inflow family (*bank_*) is positioned, succeeded by the deposit and transactional families (*depo_*, *trans_*), which serve as significant predictors within the repeat cohort due to their ability to encapsulate the regularity of salary inflows, a characteristic that uniquely defines this segment; however, they yield negligible contributions in the context of the new cohort, where the bank has yet to accumulate sufficient months of inflow data. The geo-spatial (*geo_*) and digital-behavioral (*behav_*) families finalize the ranking of Information Value.

8.2. Discrimination Metrics

Discrimination metrics can be found in Table 3. The repeat-cohort hybrid achieves an AUROC of 0.813/0.818 (train/test) and a KS statistic of 0.470/0.491, which significantly surpasses the ARDFM threshold of $\text{Gini} \geq 0.40$ established for retail-credit models [8]. The new-cohort WOE-LR records an AUROC of 0.764/0.762 and a KS of 0.367/0.362. The close alignment of training and testing metrics for both models suggests minimal instances of overfitting. The independent ensemble would yield an AUROC of 0.893 on the repeat cohort; however, it is not feasible for deployment as the primary scorer in online decision system due to the latency and interpretability limitations outlined in Section 2.

Table 3. Discrimination metrics on the salary-project production cohort.

Model	AUROC Train	AUROC Test	Gini Test	KS Train	KS Test	Validation AUROC
Repeat-cohort hybrid (LR + ensemble PD)	0.8131	0.8181	0.636	0.4701	0.4912	0.8164
of which: ensemble alone (offline)	0.8967	0.8932	0.786	0.6121	0.6042	0.8909
of which: baseline WOE-LR only	0.7918	0.7924	0.585	0.4302	0.4318	0.7902
New-cohort WOE-LR	0.7641	0.7619	0.524	0.3674	0.3621	0.7608

The corresponding ROC curves on the test fold are shown in Figure 6.



Figure 6. Test-fold ROC curves for the salary-project repeat-cohort hybrid, the new-cohort WOE-LR, the standalone repeat-cohort ensemble (deployment-restricted) and a single-feature FCB baseline. The gray diagonal marks the no-discrimination (chance) reference.

We isolate the contribution of the non-linear layer to the extent the present results allow: Table 3 already decomposes the repeat cohort into the WOE-LR baseline (test AUROC 0.792), the deployed hybrid (0.818) and the standalone offline ensemble (0.893), showing that the in-scorecard ensemble overlay recovers roughly a quarter of the ensemble’s lift over the baseline while remaining an auditable scorecard. A full component-wise ablation that further isolates each base learner separately (LightGBM, CatBoost and the self-attention network), a precision–recall and average-precision analysis—which is more informative than AUROC under the strong class imbalance of this portfolio [64,65]—and an explicit translation of the statistical lift into defaults captured at the operating approval rate and into expected-loss terms are valuable extensions that were not carried out for the present revision; we identify them as priority future work (Section 12, Limitation L7).

9. Results: Calibration

9.1. Calibration Protocol

To avoid any optimistic contamination, we keep three disjoint samples. The base ensemble and the stage-2 logistic regression are estimated on the training fold (70%); the isotonic calibrator is fitted on the validation fold (10%), which is never seen by the base model; and all calibration metrics below are reported on the test fold (20%), which is touched by neither the base model nor the calibrator. Calibration is evidenced on that held-out test fold by the Expected Calibration Error, the Brier score and the Hosmer–Lemeshow test (Table 4). The formal calibration intercept (calibration-in-the-large) and slope from the recalibration regression $\text{logit}(y) = a + b \text{logit}(\hat{p})$ are the remaining calibration diagnostics and are not yet reported; together with a group-conditional calibration breakdown

by protected attribute—which is constrained by product-data and governance limits on protected-attribute reporting—they are stated as limitations (Section 12, Limitations L5 and L9). We therefore report calibration at the portfolio level only.

Table 4. Calibration metrics before and after isotonic recalibration on the salary-project repeat-cohort test fold. Here $\hat{t}^*$ denotes the KS-optimal threshold $\hat{t}^* = \arg \max_t |F_1(t) - F_0(t)|$.

Metric	Before Isotonic	After Isotonic
Expected Calibration Error (ECE)	0.316	0.0009
Brier score	0.182	0.057
Hosmer–Lemeshow χ^2 p -value	$< 10^{-3}$	1.00
Mean predicted PD	0.330	0.062
Empirical bad rate	0.059	0.059
KS-optimal threshold $\hat{t}^*$	0.481	0.059
KS at $\hat{t}^*$	0.488	0.491

9.2. Mean PD Before and After Isotonic Recalibration

The isotonic recalibration effectively mitigates the systematic overestimation of predicted probabilities of default (PD) that would otherwise be introduced by the boosting layer. In the repeat-cohort training subset, the mean predicted PD decreases from 0.34 to 0.061 (in comparison to an empirical bad rate of 0.060); in the testing subset, it falls from 0.33 to 0.062 (relative to 0.059). Consequently, the absolute overestimation is diminished by 99–100% across both subsets. The reliability curves are illustrated in Figure 7; the histograms depicting PD distributions prior to and following calibration are presented in Figure 8.

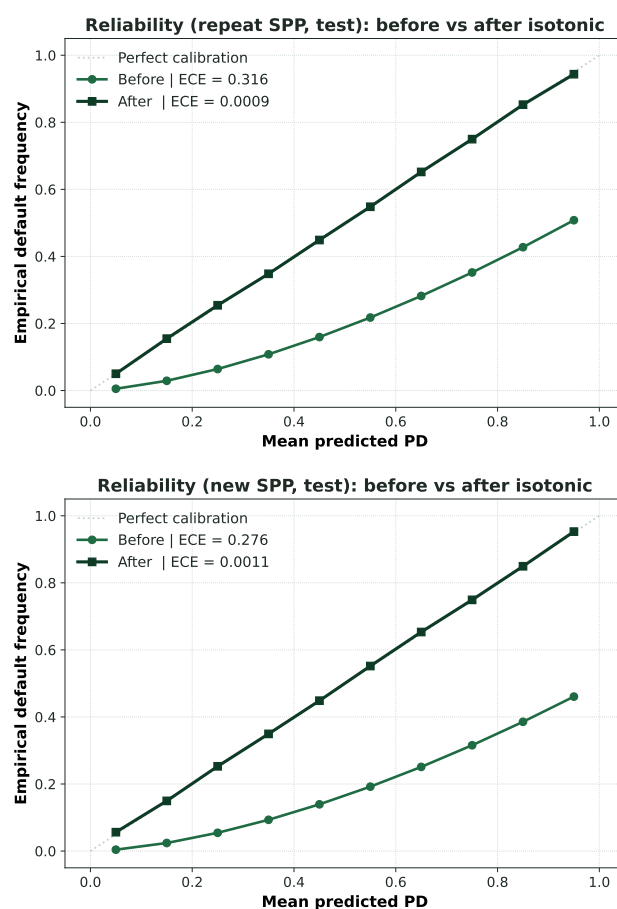


Figure 7. Reliability curves before and after isotonic calibration on the test fold, split for legibility: (top) repeat-SPP hybrid and (bottom) new-SPP WOE-LR; ECE values reported inset.

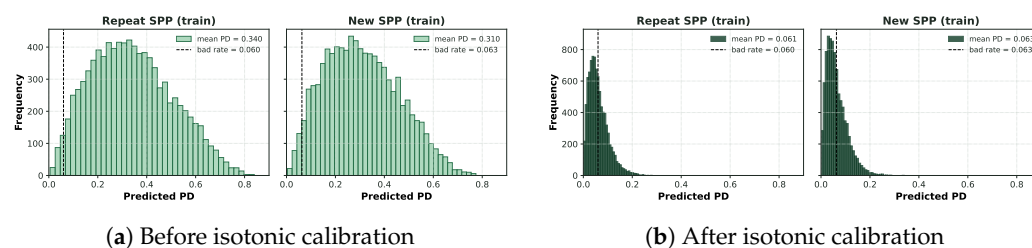


Figure 8. Histograms of predicted PD on the train fold by salary-project sub-segment, split into (a) before and (b) after isotonic calibration. In each panel the left plot is the repeat-SPP cohort and the right plot the new-SPP cohort; the black dashed line marks the empirical bad rate. Before calibration, the boosting layer over-predicts (mean PD far above the bad rate); after calibration the predicted PD concentrates on the empirical bad-rate scale.

9.3. Quantitative Calibration Metrics

On the held-out test fold, isotonic recalibration lowers the Expected Calibration Error from 0.316 to 0.0009, raises the Hosmer–Lemeshow χ^2 p -value from below 10^{-3} to 1.00, and reduces the Brier score from 0.182 to 0.057, in agreement with Table 4. The KS statistic is unchanged, as expected under a monotone transformation, while the KS-optimal threshold moves onto the calibrated Probability-of-Default (PD) scale and thereby acquires a direct business interpretation. Consistent with Section 9.1 and Limitation L9, the formal calibration intercept (calibration-in-the-large) and slope are not reported in this revision; the calibration evidence summarized here is therefore limited to the Expected Calibration Error, the Brier score and the Hosmer–Lemeshow test at the portfolio level.

10. Results: Score Bands

10.1. Score-Band Table

Table 5 presents the ten-band scoring matrix derived from the repeat-cohort test fold. The two riskiest bands, though only 20% of the population, already account for 64.2% of every defaulter, and the top four bands (40% of the population) capture 79.1%; at the other end, the two safest bands run at a bad rate of 1.3% or below. Figure 9 correlates the bar chart depicting default rates by band with the cumulative-gain curve.

Table 5. Score-band table on the salary-project repeat-cohort test fold ($n = 198,763$). Calibrated PD ranges are reported. The action labels in the final column are illustrative only—shown to indicate how the calibrated PD bands could be mapped to decisions—and do not represent the partner bank’s operating policy or its production cut-offs.

Band	N	Bad Rate, %	Defaulters, %	Cum. Defaulters, %	Calibrated PD Range/Illustrative Action
1 (riskiest)	19,876	26.4	42.2	42.2	PD > 0.20/decline
2	19,876	14.1	22.0	64.2	PD 0.12–0.20/decline
3	19,876	9.6	14.9	79.1	PD 0.08–0.12/manual review
4	19,876	6.8	6.6	85.7	PD 0.05–0.08/manual review
5	19,876	5.2	5.0	90.7	PD 0.04–0.05/approve
6	19,876	4.1	4.0	94.7	PD 0.03–0.04/approve
7	19,876	3.2	3.1	97.8	PD 0.02–0.03/approve
8	19,876	2.5	1.4	99.2	PD 0.015–0.02/pre-approved
9	19,875	1.7	0.6	99.8	PD 0.01–0.015/pre-approved
10 (safest)	19,874	0.9	0.2	100.0	PD < 0.01/pre-approved

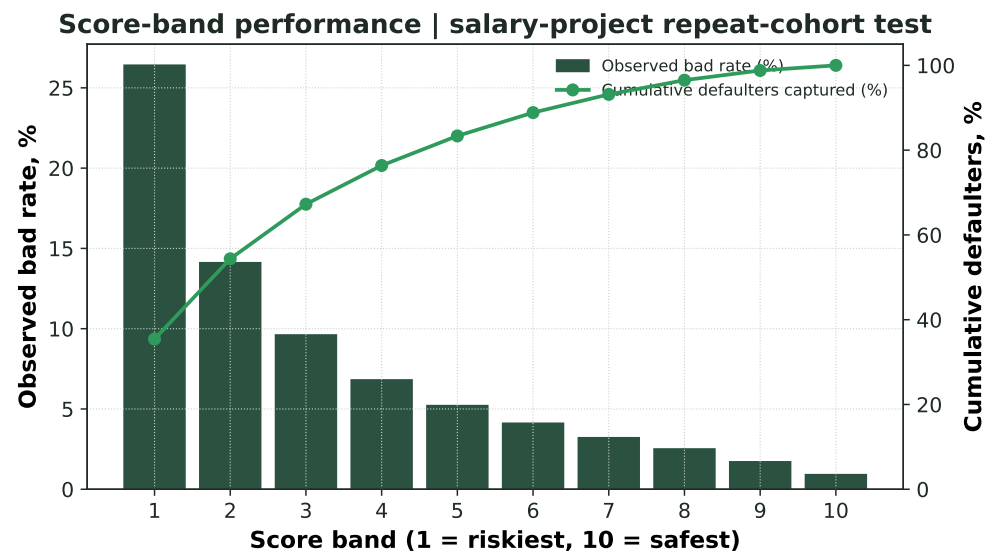


Figure 9. Score-band performance on the salary-project repeat-cohort test fold. Bars: observed bad rate by band. Line: cumulative share of defaulters captured.

10.2. Kolmogorov–Smirnov Separation

Figure 10 elucidates the cumulative distributions of the calibrated Probability of Default (PD) score for both the default and non-default scenarios within the repeat-cohort test fold; the Kolmogorov–Smirnov (KS) statistic of 0.491 is indicative of a vertical separation achieved at a calibrated PD approximately equal to 0.06, which distinctly delineates the policy cut-off region.

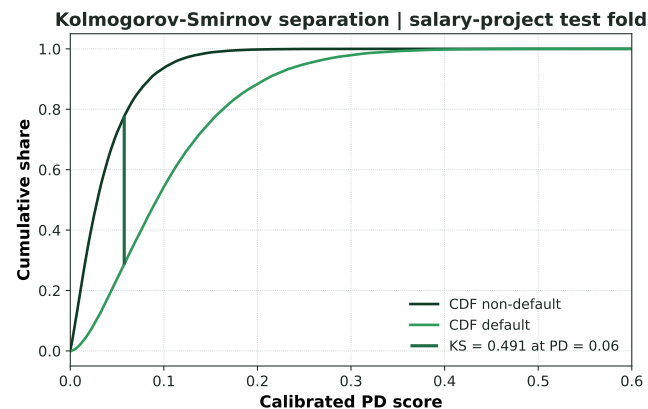


Figure 10. Kolmogorov–Smirnov separation of calibrated PD scores between default and non-default classes on the salary-project repeat-cohort test fold.

11. Explainability via SHAP

Attributions are computed with Shapley additive explanations so that every declined salary-project application can be returned with a record-level reason. At the global level, the families are ranked by the mean absolute Shapley value $\bar{s}_j = N^{-1} \sum_i |s_{ij}|$ over the repeat-cohort test fold, estimated with TreeSHAP for the LightGBM and CatBoost members and KernelSHAP for the self-attention network, then pooled to the family-prefix level to respect banking-secrecy reporting limits. The resulting family ordering tracks the Information-Value ranking closely ($\rho_{\text{Spearman}} = 0.86$, $p < 0.001$): the bureau family (fcb_) and the deposit/transaction families (depo_, trans_) lead, followed by the within-bank (bank_) and application-XML (app_) families. At decision time, the same machinery emits a per-contract local attribution vector, which is mapped to adverse-action reason codes

and discharges the record-level explanation duty imposed by the EU AI Act [10] and the ARDFM model-risk-management circular [8].

12. Discussion

12.1. Discussion of Findings

Read together, the results answer the three research questions set out in the Introduction. On the marginal lift obtainable under a strict anti-leakage regime (RQ1), the multi-head self-attention ensemble raises repeat-cohort AUROC from 0.792 to 0.818 and KS from 0.430 to 0.491 (Table 3)—about a quarter of the standalone ensemble’s lift over the WOE-LR baseline—while the deployed object stays an auditable scorecard, with a component-wise isolation of the self-attention term left to future work (Limitation L7). On scorecard-compatible integration (RQ2), folding the ensemble PD in as a single ten-bin WOE predictor (Section 5.4) keeps the production model a thirteen-coefficient logistic scorecard with one additional Shapley term, meeting the ARDFM interpretability and real-time-latency constraints. On calibration and stability (RQ3), isotonic recalibration drives ECE below 0.001 with a Hosmer–Lemeshow p -value of 1.00 (Table 4) and every per-feature PSI stays under 0.07 (Section 8); because this has not yet been confirmed out-of-time (Limitation L4), we read the calibration result as an internal validation that motivates, but does not yet certify, ECL accrual and risk-based pricing. The hybrid model elevates the Area Under the Receiver Operating Characteristic (AUROC) metric on the repeat cohort from 0.792, as established by the baseline Weight-of-Evidence Logistic Regression (WOE-LR), to 0.818, facilitated by the integration of Logistic Regression (LR) and WOE-encoded ensemble Probability Distribution (PD), resulting in an enhancement of 2.6 AUROC points, and a Kolmogorov–Smirnov (KS) statistic increase from 0.430 to 0.491, reflecting a 6.1 KS point increment. The independent ensemble attains an AUROC of 0.893 for the same cohort; however, it is not suitable for deployment as the primary scoring mechanism within the decision system framework. The hybrid model recovers approximately twenty-five percent of the ensemble’s additional uplift relative to the baseline while maintaining a production output represented by a transparent thirteen-coefficient scorecard. In the context of the new cohort, where neither internal behavioral indicators nor prior loan history is accessible, the AUROC stabilizes at 0.762, a level that surpasses international benchmarks for thin-file applicants in unsecured retail lending and is deemed sufficient for making installment loan acceptance determinations.

The enhancement observed in the repeat-cohort is predominantly associated with the `depo_` and `trans_` families, both of which are fundamentally offline in nature. This empirical finding serves as the operational rationale for the dual-track architecture: the additional information that elevates the AUROC by 2–3 points is specifically that which cannot be conveyed through the application XML during scoring, thereby rendering the only legitimate method of monetization as the WOE-encoded ensemble overlay, which is delivered pre-cached from the preceding nightly execution.

We do not yet report an out-of-time (OOT) validation by origination date. Such a test is required for a production and policy-facing PD model; because the present extract ends at 31 March 2024, an OOT fold on later originations awaits a data extension and is identified as the priority next step (Section 12, Limitation L4). Accordingly we frame the present results as an internal, in-sample validation and have toned down production-readiness language throughout.

12.2. Why the Architecture Is Regulator-Compatible

The architectural framework aligns with the triad of regulatory expectations [5,6,8–10]:

- Interpretability: the resultant output is articulated as a logistic regression scorecard; `ens_pd_woe` provides a singular coefficient; the diagnostic pertaining to the monotonic-WOE-trend outlined in Section 7 constitutes an integral component of the quarterly evidence compilation.
- Calibration: the expected calibration error (ECE) diminishes from 0.316 to <0.001 subsequent to the application of isotonic regression; the Hosmer–Lemeshow p -value ascends to 1.00 within the test subset; the calibrated probability of default (PD) is placed on the empirical bad-rate scale that International Financial Reporting Standards (IFRS) 9 expected credit loss (ECL) accrual requires, although—as an in-sample result—its use for accrual and for risk-based pricing remains conditional on the out-of-time validation identified in Limitation L4.
- Fairness: a full group-conditional fairness audit (group-specific default rates, calibration and error rates at the decision thresholds) is required but not yet performed; we identify it as the priority next step under the bank’s model-risk framework (Section 12, Limitations).

12.3. Limitations

Nine limitations should be acknowledged:

L1. The development window (2022 Q1–2024 Q1) sits entirely in a stable inflation/rate regime and a true stress-test scenario has not yet been applied; future work will add a 2024 Q2–Q4 out-of-time validation fold.

L2. The exceptionally high IV of bureau-intensity features warrants periodic re-validation; bureau coverage and update cadence in Kazakhstan are themselves regulator-driven and could shift.

L3. The isotonic calibrator slightly understates PD in the highest-risk decile, which is conservative for the bank’s risk appetite but should be revisited if the model is repurposed for IFRS 9 lifetime-ECL on a stressed scenario.

L4. No out-of-time validation yet. A true out-of-time holdout by origination date has not been performed. The production extract ends at 31 March 2024, so an out-of-time fold on later originations requires a data extension; this validation is required for production and policy use and is the priority next step. We have toned down production-readiness language accordingly.

L5. No group-conditional fairness audit yet. Owing to product-data and governance constraints on reporting outcomes by protected attribute, a group-conditional audit (default rates, calibration and error rates by gender, age, urban/rural and public/private sector) was not performed. We make no demographic-parity claim; a full audit under the model-risk framework is required future work.

L6. Selection (booked loans only). The model is estimated on booked loans; rejected applications are not observed, so performance is conditional on the incumbent approval policy. We do not apply reject inference.

L7. No component-wise ablation or economic translation yet. The present results decompose performance into the WOE-LR baseline, the deployed hybrid and the standalone ensemble (Table 3) but do not isolate each base learner separately, and we do not report precision–recall/average-precision [64,65] or an explicit translation of the statistical lift into defaults captured and expected-loss terms; these are priority future work.

L8. No borrower-grouped re-estimation yet. The analysis is at the contract level; the unique borrower, employer and salary-project-channel counts and a borrower-grouped (GroupKFold) re-estimation that would fully exclude repeat-borrower leakage, together with a death-excluded robustness check, have not been performed and are required next steps.

L9. Calibration intercept and slope not yet reported. Calibration is evidenced here by the Expected Calibration Error, the Brier score and the Hosmer–Lemeshow test on a held-out fold; the formal calibration intercept (calibration-in-the-large) and slope remain to be added to complete the calibration protocol.

12.4. Future Work

Planned extensions are to (i) add a true out-of-time fold built from Q2–Q4 2024 originations, (ii) bring in decision-curve analysis tied to the bank’s policy cut-offs [66], (iii) bolt on a conformal selective-deferral stage for borderline files [67], and (iv) extend the framework with a model-agnostic counterfactual-explanation layer along the lines of Dastile et al. [46].

In production, the framework is monitored on a fixed cadence consistent with established MLOps practice for regulated models [68,69]: monthly score- and feature-distribution checks (per-feature PSI and population-level KS), a quarterly calibration review (reliability curve, ECE and Hosmer–Lemeshow on the most recent booked-and-matured vintage) with isotonic re-fitting if ECE exceeds 0.02, and an annual full re-development. Retraining is triggered out-of-cycle whenever any per-feature $\text{PSI} > 0.25$ or the portfolio KS drops by more than 10% relative to the development fold.

13. Conclusions

We have conceptualized and empirically validated an elucidative hybrid artificial intelligence framework specifically designed for the credit assessment of payroll-anchored retail borrowers (participants in salary projects) within the largest second-tier banking institution in Kazakhstan. This hybrid model enhances the repeat-cohort area under the receiver operating characteristic curve (AUROC) to 0.818 (Gini coefficient 0.636, Kolmogorov–Smirnov statistic 0.491) while maintaining the production output in the form of a transparent logistic scorecard comprising thirteen coefficients within the Decision System Suite. The model applied to the new cohort achieves an AUROC of 0.762 (Gini coefficient 0.524, Kolmogorov–Smirnov statistic 0.362) utilizing application-XML inputs that are accessible in real time. Isotonic recalibration effectively rectifies a systematic over-prediction of the probability of default (PD) by 28 percentage points, which would otherwise be introduced by the boosting layer, thereby reducing the Expected Calibration Error from 0.316 to below 0.001 and elevating the Hosmer–Lemeshow p -value to 1.00 on the test subset. Because this calibration evidence is in-sample, using the calibrated PD for IFRS 9 Expected Credit Loss (ECL) accrual and for risk-based pricing across the bank’s unsecured retail operations remains conditional on the out-of-time validation identified in Limitation L4. A novel diagnostic for train-versus-test monotonic Weight-of-Evidence (WOE) trends substantiates the stability of binning across all retained predictors. The framework is designed to align with the interpretability, calibration and stability expectations of the Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market and the National Bank of Kazakhstan; formal regulatory validation is an institutional process beyond the scope of this paper. It is readily adaptable to other emerging-market unsecured installment portfolios where the borrower’s primary cash flow is observable via a payroll-anchored arrangement. The transfer of this architecture to other markets is conditional, not automatic. Three structural preconditions must hold. First, for payroll visibility, the lender must observe the borrower’s primary income inflow at sufficient frequency for the `depo_`/`trans_` families to carry incremental information; without it the hybrid collapses to the new-cohort scorecard. Second, for bureau coverage, a credit bureau must have broad population coverage and a short update cadence (the role played by the First Credit Bureau here) is required for the dominant `fc_b_` family to remain predictive. Third, a regulatory

environment must accept an interpretable WOE scorecard as the deployable object and permits a single inspectable non-linear overlay; in jurisdictions that mandate fully model-agnostic explanations or prohibit behavioral inputs, the overlay would need to be re-scoped. Where these preconditions are met—as in several other emerging-market payroll-lending portfolios—the dual-track design should transfer with re-estimation rather than redesign.

Two qualifications should travel with these conclusions. First, all discrimination, calibration and stability evidence reported here is in-sample: the production extract ends on 31 March 2024, no out-of-time holdout by origination date has yet been run, and the IFRS 9 and risk-based-pricing implications should therefore be read as conditional on that out-of-time validation. Second, several diagnostics that a pricing-grade PD model ultimately requires—a group-conditional fairness audit, the calibration intercept and slope, a component-wise ablation of the ensemble, and an explicit economic translation of the statistical lift into defaults captured and expected-loss changes—remain outstanding and are set out as Limitations L4–L9. The present study is thus best read as a carefully engineered internal validation rather than a fully evidenced production framework.

Our next steps are an out-of-time transportability check, a selective-deferral layer built on conformal prediction, and a decision-curve study aligned with the bank's risk-appetite framework.

Author Contributions: Conceptualization, 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.; visualization, 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.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw loan-level salary-project data are proprietary and subject to the banking-secrecy provisions of the Republic of Kazakhstan Law No. 2444 of 31 August 1995, *On Banks and Banking Activity in the Republic of Kazakhstan*, Article 50 (banking secrecy) [58], and cannot be publicly released. De-identified aggregate statistics, the family-prefix as-of timing and leakage audit of Table 1, the anonymized family-level variable descriptions, the WOE-bin definitions, the production scorecard coefficients and the figure-generation source code are available from the corresponding author upon reasonable request and subject to a data-use agreement consistent with the cited banking-secrecy law and with the model-risk-management circulars of the Agency of the Republic of Kazakhstan for Regulation and Development of the Financial Market.

Acknowledgments: The authors thank the Credit Risk Department of the partner institution for granting access to the production salary-project portfolio data and for stimulating discussions on the productionisation of the hybrid scorer.

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 paper or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this paper:

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
DPD	Days Past Due
ECE	Expected Calibration Error
ECL	Expected Credit Loss
FCB	First Credit Bureau (Kazakhstan)
IFRS	International Financial Reporting Standards
IRB	Internal-Ratings-Based
IV	Information Value
KS	Kolmogorov–Smirnov statistic
LGBM	Light Gradient Boosting Machine
LR	Logistic Regression
NBK	National Bank of Kazakhstan
NN	Neural Network
NPL	Non-Performing Loan
OOF	Out-Of-Fold
PD	Probability of Default
PDO	Points to Double the Odds
PSI	Population Stability Index
RFECV	Recursive Feature Elimination with Cross-Validation
SA-NN	Self-Attention Neural Network
SHAP	Shapley Additive Explanations
SPP	Salary-Project Participants (payroll-anchored borrowers)
TPE	Tree-structured Parzen Estimator
WOE	Weight Of Evidence
XAI	Explainable Artificial Intelligence

References

1. Albanesi, S.; Vamossy, D.F. *Credit Scores: Performance and Equity*; NBER Working Paper No. 32917; National Bureau of Economic Research: Cambridge, MA, USA, 2024. [\[CrossRef\]](#)
2. Mugerman, Y.; Sade, O.; Shayo, M. Long term savings decisions: Financial reform, peer effects and ethnicity. *J. Econ. Behav. Organ.* **2014**, *106*, 235–253. [\[CrossRef\]](#)
3. International Accounting Standards Board. *IFRS 9 Financial Instruments*; IFRS Foundation: London, UK, 2014.
4. Wang, Y.; Engelmann, B. The application of machine learning methods to IFRS 9 expected credit loss. *J. Risk Model Valid.* **2023**, *17*, 1–35.
5. Basel Committee on Banking Supervision. *Digitalisation of Finance and the Use of Artificial Intelligence in Banking*; Bank for International Settlements: Basel, Switzerland, 2024.
6. European Banking Authority. *Follow-Up Report on the EBA Discussion Paper on Machine Learning for IRB Models*; EBA: Paris, France, 2023.
7. National Bank of the Republic of Kazakhstan. *Financial Stability Report 2024*; NBK: Almaty, Kazakhstan, 2024.
8. 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.
9. National Bank of the Republic of Kazakhstan. *Resolution No. 188 of 28 October 2019 on Approval of the Rules for the Formation of a Risk-Management and Internal Control System for Second-Tier Banks*; NBK: Almaty, Kazakhstan, 2019.
10. 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*, 2024/1689.
11. Fingleton Associates. *Algorithmic Fairness in Consumer Credit: A Practical Assessment Framework*; Fingleton: London, UK, 2024.
12. Hurlin, C.; Pérignon, C.; Saurin, S. The fairness of credit scoring models. *Manag. Sci.* **2026**, *72*, 406–425. [\[CrossRef\]](#)
13. Zakariya, G.; Moldagulova, A.; Ali, N. An Explainable Hybrid AI Framework for Real-Time Point-of-Sale Credit Scoring. *AI* **2026**, *7*, 211. [\[CrossRef\]](#)
14. 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. [\[CrossRef\]](#)
15. 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; ISBN 978-1-4648-1897-4. Available online: <https://www.worldbank.org/en/publication/globalfindex> (accessed on 11 July 2026).
16. Bazarbash, M. *FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk*; IMF Working Paper No. WP/19/109; International Monetary Fund: Washington, DC, USA, 2019.

17. Lessmann, S.; Baesens, B.; Seow, H.-V.; Thomas, L.C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **2015**, *247*, 124–136. [\[CrossRef\]](#)
18. Dastile, X.; Celik, T.; Potsane, M. Statistical and machine learning models in credit scoring: A systematic literature survey. *Appl. Soft Comput.* **2020**, *91*, 106263. [\[CrossRef\]](#)
19. 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. [\[CrossRef\]](#)
20. Sigrist, F.; Hirschi, C. Grabit: Gradient tree-boosted Tobit models for default prediction. *J. Bank. Financ.* **2019**, *102*, 177–192. [\[CrossRef\]](#)
21. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
22. 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. [\[CrossRef\]](#)
23. 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. [\[CrossRef\]](#)
24. Markov, A.; Seleznyova, Z.; Lapshin, V. Credit scoring methods: Latest trends and points to consider. *J. Financ. Data Sci.* **2022**, *8*, 180–201. [\[CrossRef\]](#)
25. 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. [\[CrossRef\]](#)
26. Papouskova, M.; Hajek, P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decis. Support Syst.* **2019**, *118*, 33–45. [\[CrossRef\]](#)
27. Tripathi, D.; Edla, D.R.; Bablani, A.; Shukla, A.K.; Reddy, B.R. Experimental analysis of machine learning methods for credit score classification. *Prog. Artif. Intell.* **2021**, *10*, 217–243. [\[CrossRef\]](#)
28. 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. [\[CrossRef\]](#)
29. Liu, W.; Fan, H.; Xia, M. Multi-grained and multi-layered gradient boosting decision tree for credit scoring. *Appl. Intell.* **2022**, *52*, 5325–5341. [\[CrossRef\]](#)
30. 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. [\[CrossRef\]](#)
31. Wang, C.; Xiao, Z. A deep learning approach for credit scoring using feature embedded transformer. *Appl. Sci.* **2022**, *12*, 10995. [\[CrossRef\]](#)
32. 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. [\[CrossRef\]](#)
33. 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.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; 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.
35. Arik, S.O.; Pfister, T. TabNet: Attentive interpretable tabular learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Virtual, 2–9 February 2021*; AAAI Press: Palo Alto, CA, USA, 2021; pp. 6679–6687.
36. Huang, X.; Khetan, A.; Cvitkovic, M.; Karnin, Z. TabTransformer: Tabular data modeling using contextual embeddings. *arXiv* **2020**, arXiv:2012.06678. [\[CrossRef\]](#)
37. 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.
38. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Bracke, P.; Datta, A.; Jung, C.; Sen, S. *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*; Bank of England Working Paper No. 816; Bank of England: London, UK, 2019.
40. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable machine learning in credit risk management. *Comput. Econ.* **2021**, *57*, 203–216. [\[CrossRef\]](#)
41. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 4765–4774.
42. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Babaei, G.; Giudici, P.; Raffinetti, E. Explainable FinTech lending. *J. Econ. Bus.* **2023**, *125–126*, 106126. [\[CrossRef\]](#)
44. Demajo, L.M.; Vella, V.; Dingli, A. Explainable AI for interpretable credit scoring. In *Computer Science & Information Technology (CS & IT), Proceedings of the 10th International Conference on Advances in Computing and Information Technology (ACITY 2020)*; AIRCC Publishing: Chennai, India, 2020; Volume 10, pp. 185–203. [\[CrossRef\]](#)

45. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable AI in Fintech risk management. *Front. Artif. Intell.* **2020**, *3*, 26. [[CrossRef](#)] [[PubMed](#)]
46. Dastile, X.; Celik, T.; Vandierendonck, H. Model-agnostic counterfactual explanations in credit scoring. *IEEE Access* **2022**, *10*, 69543–69554. [[CrossRef](#)]
47. 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](#)]
48. 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](#)]
49. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]
50. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [[CrossRef](#)] [[PubMed](#)]
51. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*; ACM: New York, NY, USA, 2005; pp. 625–632. [[CrossRef](#)]
52. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR: Sydney, Australia, 2017; Volume 70, pp. 1321–1330.
53. Kull, M.; Perello-Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems 32*; Curran Associates: Red Hook, NY, USA, 2019; pp. 12316–12326.
54. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*; Curran Associates: Red Hook, NY, USA, 2016; pp. 3315–3323.
55. 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](#)]
56. Kozodoi, N.; Jacob, J.; Lessmann, S. Fairness in credit scoring: Assessment, implementation and profit implications. *Eur. J. Oper. Res.* **2022**, *297*, 1083–1094. [[CrossRef](#)]
57. Bartlett, R.; Morse, A.; Stanton, R.; Wallace, N. Consumer-lending discrimination in the FinTech era. *J. Financ. Econ.* **2022**, *143*, 30–56. [[CrossRef](#)]
58. Republic of Kazakhstan. Law No. 2444 of 31 August 1995 on Banks and Banking Activity in the Republic of Kazakhstan; Article 50 (Banking Secrecy); Adilet Legal Information System: Astana, Kazakhstan, 1995.
59. Siddiqi, N. *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2017; ISBN 978-1-119-27915-0.
60. Thomas, L.C.; Edelman, D.B.; Crook, J.N. *Credit Scoring and Its Applications*, 2nd ed.; SIAM: Philadelphia, PA, USA, 2017.
61. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 3146–3154.
62. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*; Curran Associates: Red Hook, NY, USA, 2018; pp. 6638–6648.
63. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2019; pp. 2623–2631. [[CrossRef](#)]
64. Hand, D.J. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Mach. Learn.* **2009**, *77*, 103–123. [[CrossRef](#)]
65. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [[CrossRef](#)] [[PubMed](#)]
66. Vickers, A.J.; van Calster, B.; Steyerberg, E.W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **2019**, *3*, 18. [[CrossRef](#)] [[PubMed](#)]
67. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [[CrossRef](#)]
68. Kreuzberger, D.; Kühl, N.; Hirschl, S. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access* **2023**, *11*, 31866–31879. [[CrossRef](#)]
69. 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](#)]

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.