IRSTI 28.23.15 **https://doi.org/10.58805/kazutb.v.1.30-1275**

BENCHMARKING CALIBRATED GRADIENT BOOSTING WITH BEHAVIOURAL FEATURES FOR TRANSACTION-LEVEL AML DETECTION

[1]T.K. Azbergen, [2]A.O. Tleubayeva, [1]A.T Aituov, [1]Y.T. Sabyrzhan
[1]Kazakh-British Technical University, Almaty, Kazakhstan,
[2]Astana IT University, Astana, Kazakhstan

Corresponding-author: Arailym.tll@gmail.com

Reliable transaction-level anti-money-laundering (AML) systems require not only accurate ranking of suspicious activity but also well-calibrated probabilistic risk estimates to support operational decision-making and risk prioritisation. In this context, this work provides an empirical evaluation of calibrated gradient boosting models with temporal behavioural features, focusing on the combined impact of probabilistic calibration and behavioural feature modelling on probability reliability and operationally relevant performance metrics.

The proposed framework integrates gradient boosting algorithms (CatBoost and XGBoost) with post-hoc probability calibration techniques (Platt scaling and isotonic regression) and behavioural features computed over rolling time windows of 1, 7, and 30 days to capture both short-term volatility and longer-term transaction patterns. Experimental evaluation is conducted on the publicly available IBM Anti-Money Laundering (IBM–AML) dataset, which contains approximately seven million synthetic transactions simulating realistic banking activity.

Model performance is assessed using ranking-, calibration-, and operationally oriented metrics, including AUPRC, AUROC, Brier score, Precision@Top k, and Lift@Top k. Among the evaluated baselines, the calibrated CatBoost model showed the strongest overall performance among the evaluated baselines (AUPRC = 0.367, AUROC = 0.959, Brier = 0.0062, Lift@1% = 40.9).

The results indicate that probability calibration improves the reliability of predicted risk scores without degrading ranking performance, while temporal behavioural features contribute to improved detection sensitivity in highly imbalanced settings. This study provides empirical evidence and methodological clarity regarding the role of calibration and behavioural dynamics in AML systems and offers a benchmark-ready кexperimental pipeline for future research. The applicability of the findings to real-world AML systems is discussed in light of the synthetic nature of the dataset, and directions for further validation on real banking data are outlined.

Keywords: machine learning, anti-money laundering, CatBoost, probabilistic calibration, behavioural features, financial fraud detection, explainable AI.

СРАВНИТЕЛЬНАЯ ОЦЕНКА КАЛИБРОВАННОГО ГРАДИЕНТНОГО БУСТИНГА С ПОВЕДЕНЧЕСКИМИ ПРИЗНАКАМИ ДЛЯ ВЫЯВЛЕНИЯ ОТМЫВАНИЯ ДЕНЕГ НА УРОВНЕ ТРАНЗАКЦИЙ

[1]Т.К. Азберген, [2]А.О. Тлеубаева, [1]А.Т. Айтуов, [1]Е.Т. Сабыржан
[1]Казахстанско-Британский технический университет, Алматы, Казахстан,
[2]Astana IT University, Астана, Казахстан,
e-mail: Arailym.tll@gmail.com

Надёжные системы противодействия отмыванию денег (AML) на уровне отдельных транзакций требуют не только точного ранжирования подозрительной активности, но и хорошо откалиброванных вероятностных оценок риска для поддержки операционных решений и приоритизации проверок. В этом контексте в работе представлена эмпирическая оценка калиброванных моделей градиентного бустинга с временными поведенческими признаками, направленная на анализ совместного влияния вероятностной калибровки и моделирования поведенческих признаков на надёжность вероятностных оценок и операционно значимые показатели эффективности.

Предложенная структура объединяет алгоритмы градиентного бустинга (CatBoost и XGBoost) с методами пост-калибровки вероятностей (Platt scaling и изотоническая регрессия), а также поведенческие признаки, вычисляемые в скользящих временных окнах продолжительностью 1, 7 и 30 дней, что позволяет учитывать, как краткосрочную волатильность, так и долгосрочные транзакционные паттерны. Экспериментальная оценка проведена на открытом наборе данных IBM Anti-Money Laundering (IBM–AML), содержащем около семи миллионов синтетических транзакций, имитирующих реалистичную банковскую деятельность.

Качество моделей оценивается с использованием ранжирующих, калибровочных и операционно ориентированных метрик, включая AUPRC, AUROC, Brier score, Precision@Top k и Lift@Top k. Среди рассмотренных базовых моделей калиброванная модель CatBoost продемонстрировал самые высокие общие показатели среди оцененных базовых показателей (AUPRC = 0.367, AUROC = 0.959, Brier = 0.0062, Lift@1% = 40.9).

Полученные результаты показывают, что вероятностная калибровка повышает надёжность прогнозируемых риск-оценок без ухудшения качества ранжирования, в то время как временные поведенческие признаки способствуют повышению чувствительности выявления в условиях сильного дисбаланса классов. Работа предоставляет эмпирические доказательства и методологическую ясность относительно роли калибровки и поведенческой динамики в AML-системах, а также предлагает воспроизводимую экспериментальную основу для дальнейших исследований. Применимость полученных выводов к реальным AML-системам обсуждается с учётом синтетической природы используемого набора данных, а также обозначаются направления дальнейшей валидации на реальных банковских данных.

Ключевые слова: машинное обучение, отмывание денег, CatBoost, вероятностная калибровка, поведенческие признаки, финансовое мошенничество, объяснимая ИИ.

ТРАНЗАКЦИЯЛАР ДЕҢГЕЙІНДЕ АҚШАНЫ ЖЫЛЫСТАТУДЫ АНЫҚТАУҒА АРНАЛҒАН МІНЕЗ-ҚҰЛЫҚТЫҚ БЕЛГІЛЕРГЕ НЕГІЗДЕЛГЕН КАЛИБРЛЕНГЕН ГРАДИЕНТТІ БУСТИНГТІҢ САЛЫСТЫРМАЛЫ БАҒАЛАНУЫ

[1]Т.К. Азберген, [2]А.О. Тлеубаева✉, [1]А.Т. Айтуов, [1]Е.Т. Сабыржан

[1]Қазақ-Британ техникалық университеті, Алматы, Қазақстан,

[2]Astana IT University, Астана, Казахстан,

e-mail: Arailym.tll@gmail.com

Транзакциялар деңгейіндегі ақшаны жылыстатуға қарсы (AML) сенімді жүйелер күдікті әрекеттерді дәл ранжирлеумен қатар, операциялық шешім қабылдау мен тәуекелдерді басымдықпен қарау үшін жақсы калибрленген ықтималдықтық тәуекел бағаларын талап етеді. Осы тұрғыда бұл жұмыста уақыттық мінез-құлықтық белгілерді қолданатын калибрленген градиентті бустинг модельдерінің эмпирикалық бағалануы ұсынылады және ықтималдық калибрлеу мен мінез-құлықтық белгілерді модельдеудің ықтималдық бағалардың сенімділігіне және операциялық тұрғыдан маңызды тиімділік көрсеткіштеріне бірлескен әсері талданады.

Ұсынылған құрылым градиентті бустинг алгоритмдерін (CatBoost және XGBoost) ықтималдықтарды пост-калибрлеу әдістерімен (Platt scaling және изотондық регрессия) біріктіреді, сондай-ақ 1, 7 және 30 күндік жылжымалы уақыттық терезелерде есептелетін мінез-құлықтық белгілерді пайдаланады. Бұл тәсіл қысқа мерзімді құбылмалылықты да, ұзақ мерзімді транзакциялық үлгілерді де қамтуға мүмкіндік береді. Эксперименттік бағалау нақты банктік қызметті модельдейтін шамамен жеті миллион синтетикалық транзакциядан тұратын IBM Anti-Money Laundering (IBM–AML) ашық деректер жиынтығында жүргізілді.

Модельдердің тиімділігі ранжирлеу, калибрлеу және операциялық бағытталған метрикалар арқылы бағаланды, олардың қатарына AUPRC, AUROC, Brier көрсеткіші, Precision@Top k және

Lift@Top k жатады. Қарастырылған базалық модельдер арасында калибрленген CatBoost бағаланған базалық көрсеткіштер арасында ең жоғары жалпы көрсеткіштерді көрсетті (AUPRC = 0.367, AUROC = 0.959, Brier = 0.0062, Lift@1% = 40.9).

Нәтижелер ықтималдық калибрлеудің модельдің ранжирлеу қабілетін төмендетпей, болжанған тәуекел бағаларының сенімділігін арттыратынын, ал уақыттық мінез-құлықтық белгілердің сыныптар қатты теңгерімсіз болған жағдайда анықтау сезімталдығын күшейтетінін көрсетеді. Зерттеу AML жүйелеріндегі калибрлеу мен мінез-құлықтық динамиканың рөліне қатысты эмпирикалық дәлелдер мен әдістемелік айқындықты ұсынады және болашақ зерттеулер үшін қайта өндіруге болатын эксперименттік негіз қалыптастырады. Алынған нәтижелердің нақты AML жүйелеріне қолданылу мүмкіндігі пайдаланылған деректер жиынтығының синтетикалық сипатын ескере отырып талқыланып, нақты банктік деректерде қосымша валидация жүргізу бағыттары айқындалады.

Түйін сөздер: машиналық оқыту, ақшаны жылыстату, CatBoost, ықтималдық калибрлеу, мінез-құлықтық ерекшеліктер, қаржылық алаяқтық, түсіндірілетін жасанды интеллект.

Introduction. Money laundering (AML) remains one of the most persistent challenges to global financial stability, affecting both the integrity of banking institutions and the credibility of regulatory systems. According to the Financial Action Task Force (FATF), illicit financial flows represent approximately 2-5% of global GDP, illustrating the magnitude of the threat and the necessity for advanced analytical tools to counteract it. The rapid evolution of laundering schemes, the expansion of digital banking platforms, and the increasing complexity of transactional ecosystems have made the detection of suspicious activity a pressing global priority [1].

Traditional rule-based AML systems, which have long dominated industry practice, rely on static thresholds, expert-defined rules, and manually specified behavioural heuristics [2]. While these systems are transparent and straightforward to audit, they exhibit several critical limitations. They tend to produce excessively high false-positive rates-often exceeding 90% of all alerts [3], which leads to a heavy compliance burden, inflated operational costs, and reduced confidence in automated systems. Moreover, rule-based models struggle to adapt to new or evolving laundering typologies, as their decision logic remains fixed once deployed.

In contrast, machine learning (ML) offers a data-driven and adaptive alternative capable of identifying complex behavioural and temporal patterns in financial transactions [4]. ML-based approaches enable dynamic adjustment to new data distributions, capture non-linear relationships, and provide improved detection performance across heterogeneous client profiles and transaction types [5]. This shift reflects a broader paradigm in financial monitoring: from deterministic rules toward statistically grounded, probabilistically interpretable models that learn from empirical evidence.

Recent studies have highlighted the effectiveness of supervised ML models in AML detection, frequently reporting that gradient boosting methods-such as XGBoost, LightGBM, and CatBoost-outperform traditional rule-based heuristics in identifying suspicious transactions. However, reported performance gains are often highly sensitive to feature construction strategies, data-splitting protocols, and the choice of evaluation metrics. In particular, much of the existing literature prioritises global discrimination measures, such as the area under the receiver operating characteristic curve (ROC-AUC) and the F1-score, while giving comparatively limited attention to probability reliability and operational prioritization [6].

As a result, despite achieving strong ranking performance, many AML models produce risk scores whose probabilistic interpretation is unclear or unreliable, limiting their usefulness for risk-based decision-making in compliance workflows. Two aspects are especially critical in this regard. First, probability calibration ensures that predicted risk scores correspond to empirically valid probabilities, enabling consistent prioritisation of

alerts under constrained review resources. Second, temporal behavioural representation captures the dynamic evolution of client activity over time, rather than relying on static transaction snapshots, thereby improving sensitivity to subtle and emerging laundering patterns.

In parallel, graph-based learning approaches have been explored to represent financial systems as networks of interconnected accounts, entities, and transactions. Graph Neural Networks (GNNs) [7], including Graph Convolutional Networks (GCNs) [8], dynamic GNN architectures such as MDGC-LSTM [9], and multi-view formulations such as LineMVGNN [10], have demonstrated promise in modelling multi-stage laundering schemes and uncovering hidden relational dependencies. Nevertheless, many graph-based AML models rely on uncalibrated outputs and provide limited temporal context, which can constrain their interpretability, stability, and suitability for operational compliance settings.

Compared to prior work, existing behavioural ML approaches [6] primarily emphasise feature design and anomaly sensitivity, but typically evaluate models using discrimination-focused metrics without explicitly addressing probability calibration or operational prioritisation. At the same time, recent graph-based AML studies [7-10] demonstrate the value of relational representations for uncovering complex laundering structures, yet often rely on uncalibrated outputs and lack systematic temporal aggregation aligned with real monitoring workflows. As a result, probability reliability and risk score interpretability-both critical for compliance decision-making-remain underexplored in much of the existing literature.

Taken together, these observations highlight a clear research gap. Existing AML models rarely combine temporal behavioural modelling, probability calibration, and graph-based contextual learning within a unified, reproducible framework. Consequently, this study aims to empirically investigate this gap by evaluating how calibration-aware and behaviour-aware machine-learning pipelines perform under operationally relevant AML metrics, rather than by proposing a new detection paradigm.

Accordingly, this work is positioned as an empirical evaluation study. The main contribution is an empirically grounded evaluation protocol that jointly analyses calibration, behavioural temporal features, and operational top-k metrics within a reproducible benchmark setting.

The object of this research is the process of detecting and classifying suspicious financial transactions within an AML system. The subject of the research is the application of calibrated and behaviour-aware machine-learning methods for improving the reliability of AML risk scoring. The purpose of the study is to empirically assess an integrated AML framework that combines temporal behavioural features and probabilistic calibration under ranking- and operation-oriented evaluation criteria.

This work is based on the hypothesis that the integration of temporal behavioural features and probability calibration can improve ranking stability and probability reliability in highly imbalanced AML detection tasks. Specifically, calibrated boosting algorithms such as CatBoost, when combined with rolling behavioural features, can yield more stable and better-calibrated AML risk scores than uncalibrated or purely graph-based models.

Accordingly, the study pursues the following objectives:

To critically analyse the limitations of rule-based and conventional ML approaches in AML detection.

- To design a behavioural feature-engineering pipeline that captures temporal transaction dynamics through rolling windows of 1, 7, and 30 days.

- To implement and compare probability calibration techniques, including Platt scaling and isotonic regression.

- To evaluate and contrast the performance of calibrated boosting models (CatBoost, XGBoost) and graph-based architectures (EdgeSAGE) on a unified dataset.

- To develop a reproducible experimental framework based on the IBM-AML dataset, enabling transparent benchmarking for future research.

The methodological framework combines supervised learning, probability calibration, and graph-based analysis. Feature engineering incorporates rolling behavioural statistics to capture temporal variability in transaction activity.

The dataset employed - the IBM Anti-Money Laundering (AML) dataset provides a realistic synthetic simulation of financial transactions [11]. Model performance is assessed using ranking metrics (AUPRC, AUROC) [12], calibration metrics (Brier score, reliability curves) [13], and operational metrics (Precision@Top k, Lift@Top k) [14]. All experiments follow a temporal data-splitting strategy to preserve chronological order and simulate realistic monitoring scenarios.

The scientific novelty of this study does not consist in the development of a fundamentally new machine-learning algorithm. Instead, it lies in a structured and reproducible empirical analysis of how temporal behavioural modelling and probabilistic calibration jointly affect reliability and operational interpretability in AML detection tasks.

Unlike prior research focused primarily on discrimination accuracy, this work emphasises probability reliability and operational evaluation metrics as complementary dimensions of AML model assessment. This focus is motivated by the practical need to support risk prioritisation and auditability, where relative ranking alone is insufficient without reliable probability estimates.

The theoretical significance of the study lies in advancing empirical understanding of calibration-aware and behaviour-aware machine-learning approaches in financial compliance.

The practical significance is reflected in a benchmark-ready evaluation pipeline that can support future comparative studies and validation on real-world financial data.

Materials and methods. The experiments were conducted using the IBM Anti-Money Laundering (AML) dataset, a publicly available synthetic dataset designed to emulate real-world financial transactions. It contains approximately 7 million transaction records, representing a virtual banking environment with both legitimate and illicit financial activities. Each transaction includes structured attributes describing the sender, receiver, transaction amount, currency, payment type, and temporal information.

Since the original dataset is highly imbalanced, with an approximate ratio of 1:200 between suspicious and non-suspicious transactions, a stratified sampling procedure was applied prior to model training to reduce the dataset size while preserving the original class distribution. Temporal ordering was maintained during the train-validation-test split to prevent information leakage, resulting in a manageable and statistically representative subset of approximately 1.5 million records suitable for training and evaluation.

To maintain temporal consistency and reflect real monitoring conditions, a chronological data split was employed:

- 70% of transactions for training,

- 15% for validation, and

- 15% for testing.

This temporal partition prevents information leakage and allows for backtesting across monthly periods (e.g., 2022-09) to assess model stability over time.

Feature engineering is central to this study and aims to capture behavioural and temporal patterns in transaction dynamics.

Three major feature categories were developed:

1. Categorical features - nominal attributes describing transaction type, currency, bank identifier, and payment format (≈5 variables). These were encoded using target or one-hot encoding depending on cardinality.

2. Quantitative features - numerical indicators such as transaction amount, average, standard deviation, and count of transactions.

3. Behavioural rolling features - temporal aggregates computed over 1-, 7-, and 30-day sliding

windows, including:

- mean and median transaction values;

- transaction count per window;

- rolling sum and deviation (MAD, standard deviation) [15].

These features represent short-term and long-term behavioural trends, allowing the model to detect deviations from normal patterns that may indicate laundering activities.
All numerical features were scaled using z-score normalization, and missing values were handled through median imputation.

The proposed AML detection framework integrates both classical machine-learning models and graph-based architectures, enabling comparative evaluation of their predictive and operational effectiveness.

Four supervised learning algorithms were trained on the tabular feature set:

1. Logistic Regression (LR) - interpretable linear baseline;

2. Random Forest (RF) - ensemble of decision trees for non-linear modelling;

3. XGBoost - gradient boosting framework with efficient regularization;

4. CatBoost - categorical boosting model optimized for mixed data types.

CatBoost and XGBoost were selected as primary baselines because gradient boosting is a widely adopted and highly effective model class for heterogeneous tabular data, including imbalanced classification settings [16, 17]. CatBoost additionally provides built-in mechanisms for handling categorical features and was designed to reduce prediction shift caused by target leakage during boosting, which is beneficial in mixed-type transaction data [18].

We focused on CatBoost and XGBoost as representative, widely deployed gradient-boosting implementations with mature open-source ecosystems and strong empirical performance on heterogeneous tabular data. LightGBM was not included to keep the benchmark compact and avoid overlapping baselines with similar inductive bias under the same feature set; its inclusion is left for future extensions of the benchmark. Neural tabular models were deprioritised due to higher optimisation complexity and stronger dependence on tuning choices, which can increase training variance in operational pipelines.

Alternative deep neural models were not prioritised due to higher optimisation complexity and stronger dependence on training and tuning choices, which can increase variance and computational cost in practical tabular pipelines [19, 20]. Graph neural networks were included as a conceptual baseline for relational learning over account–transaction graphs, following standard inductive graph representation learning frameworks [21, 22].

Each model was trained using a binary classification objective (isMoneyLaundering = 1 vs.0), with class weights adjusted to counter data imbalance.

Hyperparameters were optimized on the validation set via grid search, with early stopping based on validation AUPRC. To reduce sensitivity to hyperparameter choice, all models were evaluated using fixed validation-based configurations, and performance stability was assessed through temporal backtesting and feature ablation experiments.

To assess robustness with respect to hyperparameter choice, the following tuning policy was adopted. Hyperparameter search was intentionally constrained to a compact grid to reduce variance and avoid overfitting to a single validation period, in line with established recommendations for ensemble models [23]. Gradient boosting methods are known to exhibit relatively low sensitivity to moderate variations in depth and learning rate due to their ensemble structure [17]. In our experiments, CatBoost and XGBoost demonstrated stable performance across reasonable hyperparameter ranges, with only minor fluctuations in AUPRC and top-k operational metrics, while post-hoc probability calibration consistently improved the Brier score, as expected from prior calibration studies [17,24].

To enhance interpretability and improve risk-based ranking, model outputs were calibrated using two standard post-processing techniques:

1. Platt Scaling (Sigmoid Calibration) - fits a logistic regression model on validation predictions [25];

2. Isotonic Regression - non-parametric monotonic mapping to align predicted probabilities with empirical frequencies [24].

3. Calibration models were fitted on the validation set and applied to test predictions to ensure unbiased evaluation.

Calibration was applied to improve probability reliability and was evaluated using the Brier score and reliability diagrams in the Results section.

In addition to tabular ML models, a graph-based EdgeSAGE model was implemented. The transaction network was constructed by connecting nodes representing accounts through edges representing transactions [22].

Each edge carried feature vectors derived from the transaction attributes. The EdgeSAGE model aggregated information from neighbouring nodes to predict the likelihood of a transaction being suspicious.

The EdgeSAGE model was included as a graph-based baseline to enable comparative evaluation with tabular models under the same experimental conditions.

Model performance was evaluated using three complementary groups of metrics designed to capture ranking quality, probabilistic calibration, and operational effectiveness.

Given the strong class imbalance of the AML dataset (approximately 1:200), ranking and precision-based metrics were prioritised over accuracy-based measures to provide a more realistic assessment of detection capability.

Model calibration was performed using Platt scaling [25] and isotonic regression [24], fitted on validation predictions.

The AUPRC quantifies the trade-off between precision and recall across varying decision thresholds, making it particularly suitable for highly imbalanced datasets such as AML detection.

Precision (P) and Recall (R) are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

where:

- TP - number of true positives (correctly identified suspicious transactions),

- FP - number of false positives,

- FN - number of false negatives.

The area under the curve (AUPRC) is then computed as the integral of precision with respect to recall:

$$\text{AUPRC} = \int_0^1 P(R)\, dR \tag{2}$$

Higher AUROC indicates better ranking of positive (suspicious) instances among top risk scores.

The AUROC evaluates the model's ability to discriminate between positive and negative classes across all thresholds. It is defined as the probability that a randomly chosen positive instance receives a higher predicted score than a randomly chosen negative instance:

$$\text{AUROC} = \int_0^1 TPR(FPR)\, dFPR \tag{3}$$

where:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{4}$$

and TN denotes true negatives. While AUROC is a standard indicator of model discrimination, AUPRC is generally more informative for imbalanced AML data where false negatives are critical.

The Brier score (BS) measures the mean squared difference between predicted probabilities (p_i) and actual outcomes (y_i):

$$\text{Brier score} = \frac{1}{N} i = 1N \left(\widehat{p}_i - y_i\right)^2 \tag{5}$$

where:

- $p_i \in [0,1]$ is the predicted probability of laundering for transaction i,

- $y_i \in \{0,1\}$ is the ground-truth label,

- N is the number of samples

A lower Brier score indicates better probabilistic reliability and calibration of the model.

To visually assess calibration, a reliability (calibration) curve plots the average predicted probability against the observed frequency of positive cases in corresponding probability bins.

Perfect calibration corresponds to the diagonal y=x, y=x.

Deviations from this line indicate overconfidence (predicted probabilities too high) or underconfidence (too low). In AML contexts, well-calibrated risk scores are essential for prioritising transactions according to realistic probability estimates rather than arbitrary ranking.

To reflect resource-constrained review settings, models are evaluated under top-k selection scenarios, where only the highest-ranked transactions are examined.

Precision@Topk measures the proportion of true positives within the top $k\%$ of transactions with the highest predicted scores:

$$\text{PrecisionTopk} = TP_k TPk + FPk \qquad (6)$$

where TPk and FPk denote true and false positives among the top $k\%$ ranked cases.

Recall@Top k measures how many of the actual laundering cases are captured in the top $k\%$ predictions:

$$\text{RecallTopk} = TP_k TP \qquad (7)$$

This metric reflects the system's ability to capture critical alerts within the operational review budget (e.g., the top 1% of transactions reviewed daily by analysts).

The Lift metric quantifies how much better the model performs than random selection at a given cutoff. It is computed as the ratio between the model's Precision@Top k and the base rate of positives in the dataset:

$$LiftatTopk = \frac{\text{Precision at Top } k}{P_{\text{base}}}, \quad P_{\text{base}} = \frac{TP + FP}{N} \qquad (8)$$

Higher Lift@k values indicate a greater concentration of true positive cases within the top-ranked transactions compared to random selection. In applied AML evaluation scenarios, Lift@1% values above 30× are commonly interpreted as indicating a high concentration of detected cases within a limited review subset.

Results and discussion. The experimental evaluation was conducted on the IBM-AML dataset [11], comprising several million synthetic financial transactions that simulate realistic banking behaviour, including both legitimate and laundering operations.

The performance of multiple machine-learning models -Logistic Regression, Random Forest, XGBoost, CatBoost, and a Graph Neural Network (EdgeSAGE) - was assessed using three complementary groups of metrics: ranking, calibration, and operational effectiveness.

Table 1 summarises the comparative performance of all baseline models evaluated on the IBM-AML dataset.

In contrast to prior studies focusing primarily on ROC-AUC or F1-score optimisation [3,5], the present results indicate that calibration-aware models can improve probability reliability without sacrificing ranking performance.

This observation complements behavioural feature-based approaches [6] by quantifying calibration effects under operational AML metrics.

The CatBoost model with Platt calibration achieved the strongest overall performance among the evaluated models under the selected ranking, calibration, and operational metrics, with an AUPRC of 0.367 and Lift@1% of 40.9×, indicating that a large proportion of true laundering cases is concentrated within the top-ranked transactions.

Table 1 - Performance comparison of baseline models on the IBM-AML dataset

Model	AUPRC	AUROC	Brier	Precision@1%	Recall@1%	Lift@1%
Logistic Regression	0.0479	0.894	0.0067	0.069	0.099	9.9×
Random Forest	0.0904	0.913	0.0066	0.153	0.219	21.9×
XGBoost (cal.)	0.1680	0.946	0.0062	0.196	0.282	28.2×
CatBoost (cal.)	**0.3674**	**0.959**	**0.0062**	**0.285**	**0.409**	**40.9×**
Graph Neural Network (EdgeSAGE)	0.0110	0.666	0.772	0.007	0.009	0.9×

This corresponds to an approximately four-fold improvement in operational precision relative to XGBoost and a substantial gain over random selection under the chosen top-k operational setting. The Brier score of 0.0062 indicates strong probabilistic calibration, reflecting improved alignment between predicted risk scores and observed event frequencies. These estimates are specific to the IBM-AML benchmark and may not directly transfer to real-world deployments.

In contrast, the Graph Neural Network (EdgeSAGE) underperformed substantially (AUPRC = 0.011, AUROC = 0.666), primarily due to the limited depth of neighbourhood aggregation and absence of temporal contextualisation.

This result highlights the critical role of behavioural and temporal dynamics in AML detection, which conventional GNNs, when applied without explicit temporal modelling, may not adequately capture. Table 2 presents a focused comparison of CatBoost models before and after probability calibration

Table 2 - Effect of probability calibration on CatBoost performance

Model	AUPRC	Brier	Precision@1%	Recall@1%	Lift@1%
Uncalibrated CatBoost	0.3674	0.1031	0.285	0.409	40.9×
Platt (Sigmoid) Calibration	**0.3674**	**0.0062**	**0.285**	**0.409**	**40.9×**
Isotonic Calibration	0.3439	**0.0054**	0.281	0.403	40.4×

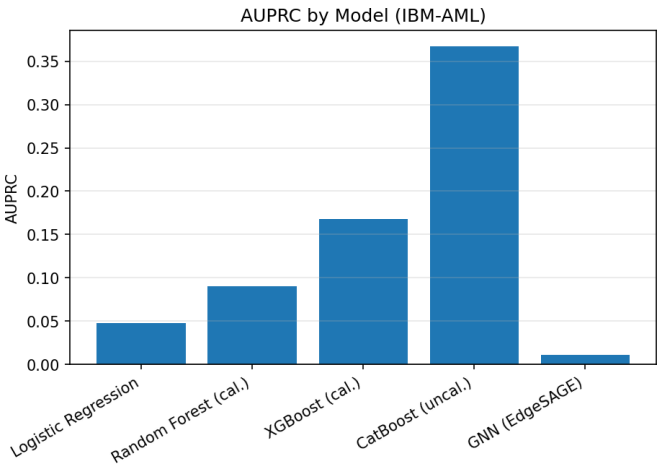


Fig.1 - AUPRC by model on the IBM-AML dataset

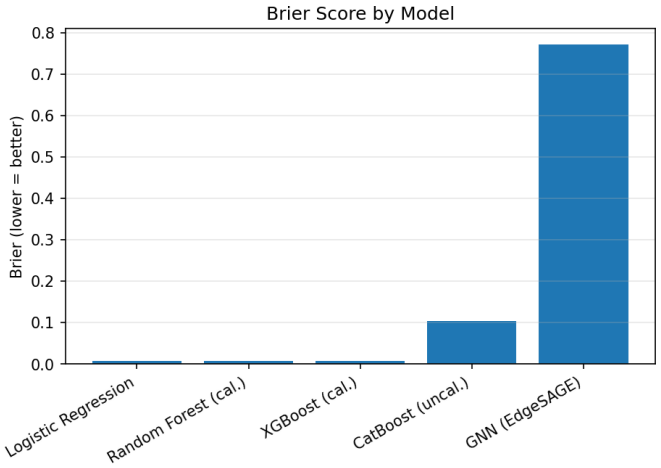


Fig.2 - Brier Score by model

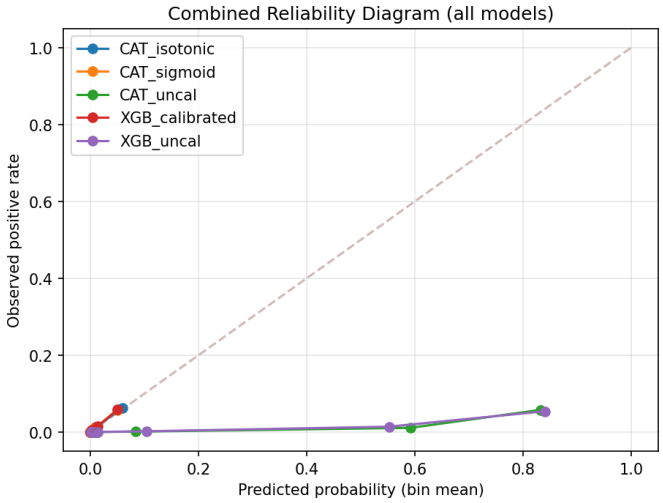


Fig.3 - Combined reliability diagram for all models

Both Platt scaling and isotonic regression markedly improved probability reliability without reducing ranking accuracy.

Although AUPRC remained nearly constant, the Brier score decreased by nearly **17×**, confirming a major enhancement in the model's probabilistic consistency.

Between the two techniques, isotonic regression achieved marginally better local calibration (lower Brier) but incurred a slight global ranking penalty (AUPRC ↓ 0.024).

Figure 1 shows that CatBoost outperforms all other models in AUPRC. Figure 2 compares Brier scores, confirming superior calibration, while Figure 3 presents reliability curves indicating near-linear alignment between predicted and empirical probabilities, suggesting improved probability interpretability.

To analyse the impact of temporal and dispersion-based behavioural features, an ablation study was performed by sequentially excluding specific feature groups. Table 3 summarises the results.

Table 3 - Ablation results for behavioural features

Removed Feature Group	ΔAUPRC	ΔLift@1%	Interpretation
1-Day Rolling Window	−0.012	−2.0×	Short-term volatility features improve responsiveness to rapid behaviour changes.
MAD (Mean Absolute Deviation)	−0.008	−1.5×	Dispersion-based features enhance anomaly sensitivity in low-signal regions.
Both (1-Day + MAD)	**−0.023**	**−4.0×**	Combined removal leads to noticeable degradation in ranking precision.

These results indicate that short-term (1-day) and dispersion (MAD) features make a substantial contribution to AML detection sensitivity, particularly in rare-event contexts.

Their removal caused a measurable degradation in both AUPRC and Lift@1 %, indicating that behavioural temporal dynamics play an important role in robust detection.

A backtesting experiment was conducted on the September 2022 period to evaluate model robustness across time. To assess the temporal robustness of the proposed models, a backtesting

analysis was conducted on the September 2022 transaction period. The results are presented in Table 4 and visualised in Figure 4.

Table 4 - Monthly backtesting results (AUPRC) on the IBM-AML dataset

Month	AUPRC (XGB)	AUPRC (CAT)	n (Transactions)
2022-09	0.1784	0.3674	107 484

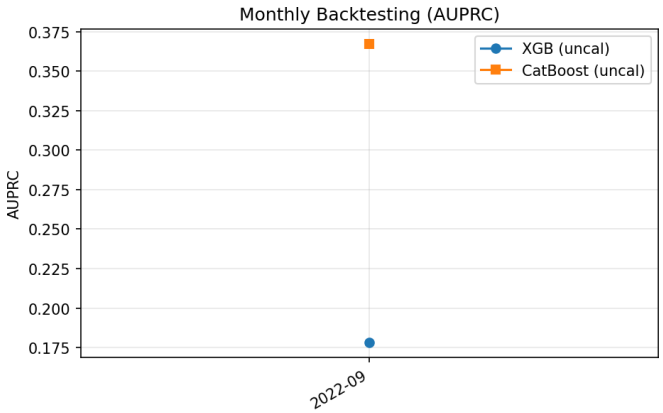


Fig. 4 - Temporal backtesting results (AUPRC): comparison of CatBoost and XGBoost models on the September 2022 transaction period

CatBoost maintained stable AUPRC ≈ 0.37 across time, outperforming XGBoost (0.18). This temporal robustness indicates resilience to distribution shifts, an important consideration for operational AML settings.

However, analysis revealed clusters of false positives among accounts exhibiting short-term transaction bursts (e.g., payroll disbursements or high-frequency business activity). These legitimate but transient patterns can be misclassified as laundering. Future work will incorporate temporal decay factors and entity-type embeddings to distinguish transient anomalies from systematic laundering behaviour.

The combined analysis supports the hypothesis that calibrated, behaviour-aware models outperform both uncalibrated and purely graph-based approaches. Behavioural rolling features enhance sensitivity to dynamic patterns, while probability calibration ensures reliability and interpretability of risk scores.

From an operational perspective, the calibrated CatBoost model concentrates a substantially higher proportion of true positives within the top-ranked transactions, which may support more efficient prioritisation in AML review processes.

At the same time, these results should be interpreted as indicative, as operational deployment would require additional validation, integration, and regulatory assessment.

Nonetheless, the use of a synthetic dataset (IBM-AML) constitutes a limitation, as it may not fully reflect the diversity and noise of real-world financial data. Future research will expand this framework using multi-institutional datasets, temporal graph neural networks (TGNNs), and explainability methods (SHAP, LIME) to enhance interpretability and cross-domain generalisation.

In summary, the results indicate that integrating temporal behavioural modelling with probabilistic calibration improves ranking stability and probability reliability in transaction-level AML detection. Behavioural rolling features enhance sensitivity to dynamic patterns, while post-hoc

calibration improves the consistency of predicted risk scores.

At the same time, the reliance on a synthetic dataset (IBM-AML) constitutes a limitation, and the reported findings should be interpreted as indicative rather than directly transferable to production environments. Further validation on real-world banking data, along with the incorporation of temporal graph-based models and explainability techniques, is necessary to assess practical applicability.

A key limitations of this study is the reliance on a synthetic dataset (IBM–AML), which, although widely used in academic research, may not fully capture the heterogeneity, noise, and adaptive behaviour observed in real-world financial systems.

As a result, the reported performance metrics should be interpreted as indicative rather than directly transferable to production AML environments.

Future work will therefore focus on validation using real or quasi-real multi-institutional datasets.

Ethical considerations. This study uses a publicly available synthetic dataset and does not involve personal or confidential financial records. Nevertheless, automated AML decision support may introduce risks related to false positives, biased prioritisation, and accountability; therefore, future work on real-world deployment will include fairness, explainability, and governance-oriented evaluation.

Conclusion. This study presented a reproducible machine-learning framework for transaction-level anti-money-laundering (AML) detection that integrates temporal behavioural feature modelling with probabilistic calibration. Experimental evaluation on the IBM Anti-Money Laundering (IBM-AML) dataset demonstrated that calibrated gradient boosting models, particularly CatBoost, achieve strong ranking performance while improving probability, reliability, as reflected by a low Brier score and stable operational metrics.

The results provide empirical evidence that the combination of rolling behavioural features and post-hoc calibration enhances the robustness and interpretability of AML risk scoring without degrading discriminative performance. Short-term and dispersion-based behavioural features were shown to be especially important in highly imbalanced detection settings, while Platt scaling and isotonic regression effectively aligned predicted scores with observed outcome frequencies.

From a methodological perspective, the study contributes an integrative evaluation framework that jointly considers ranking quality, calibration, and operational effectiveness, which are often analysed separately in prior AML research. From a practical standpoint, the findings highlight the potential value of calibration-aware and behaviour-aware models for risk prioritisation in financial monitoring systems.

At the same time, the reliance on a synthetic dataset constitutes a limitation, and the reported results should be interpreted as indicative rather than directly transferable to production environments. Future work will focus on validation using real-world banking data, incorporation of temporal graph-based models, and the application of explainability techniques to further support transparent and auditable AML decision-making.

References

1. Goldbarsht D., Benson K. From later to sooner: exploring compliance with the global regime of anti-money laundering and counter-terrorist financing in the legal profession // Journal of Financial Crime. -2023.-Vol.31(4). - P.795-809. DOI 10.1108/jfc-08-2023-0201.

2. Utama P.I., Novianto W.T., Purwadi H. The role of notaries/land deed officials in combating money laundering crimes with the know your customer principle // International Journal of Educational Research & Social Sciences. -2024.-Vol.5(2). - P.279-283. DOI 10.51601/ijersc. v5i2.801.

3. Reite E., Karlsen J., Westgaard E. Improving client risk classification with machine learning to increase anti-money laundering detection efficiency // Journal of Money Laundering Control. -2025.-Vol.28(1). P.93-107.DOI 10.1108/JMLC-03-2024-0040.

4. Ali A., Abd Razak S., Othman S.H., Eisa T.A.E., Al-Dhaqm A., Nasser M., Elhassan T., Elshafie H., Saif A. Financial fraud detection based on machine learning: a systematic literature review // Applied Sciences. 2022.- Vol.12(19). - P.1-24. DOI 10.3390/app12199637.

5. Juneja S., Saurav A., Prabhakar A., Anand V., Goyal R., Thakur V. Enhancing fraud detection with gradient boosting algorithms in credit card transactions//Proc. of the 2025 Int. Conf. on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI). -2025.-P.1-8. DOI 10.1109/IC3ECSBHI63591.2025.10990658.

6. Lopata A., Gudas S., Butleris R., Rudžionis V., Žioba L., Veitaitė I., Dilijonas D., Grišius E., Zwitserloot M. Financial data anomaly discovery using behavioral change indicators // Electronics. - 2022. Vol.11(10). -P.1-14. DOI 10.3390/electronics11101598.

7. Weber M., Domeniconi G., Chen J., Weidele D.K.I., Bellei C., Robinson T., Leiserson C.E. Anti-money laundering in bitcoin: experimenting with graph convolutional networks for financial forensics//arXiv preprint.-2019.-P.1-7.DOI 10.48550/arXiv.1908.02591.

8. Johannessen F., Jullum M. Finding money launderers using heterogeneous graph neural networks//arXiv preprint. -2023.-P.1-20. DOI: 10.48550/arXiv.2307.13499.

9. Wan F., Li P. A novel money laundering prediction model based on a dynamic graph convolutional neural network and long short-term memory//Symmetry. -2024.-Vol.16.-P.1-22. DOI 10.3390/sym16030378.

10. Poon C.-H., Kwok J., Chow C., Choi J.-H. LineMVGNN: anti-money laundering with line-graph-assisted multi-view graph neural networks //AI. -2025.-Vol.6(4). DOI 10.3390/ai6040069.

11. IBM/AML-Data. IBM Transactions for Anti-Money Laundering (AML) Dataset [Online resource].-2019.URL:https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml.- Date of access: 10.10.2025.

12. Adelino J.M., Tierra M.C.M., Villafuerte M.Q. Development and evaluation of multi-week tropical cyclone strike probability forecasts in the Philippines // Philippine Journal of Science. -2023.-Vol.152(SI). -P.1-13.DOI 10.56899/152.s1.09.

13. Xu P., Davoine F., Zha H., Denœux T. Evidential calibration of binary SVM classifiers //Approximate Reasoning. -2016.-Vol.-72.-P.55-70. DOI 10.1016/j.ijar.2015.05.002.

14. Nguyen H.D., Fryer D. A binary-response regression model based on support vector machines // arXiv preprint. -2020. P.1-20. DOI 10.48550/arXiv.2007.11902.

15. Silva Filho T., Song H., Perello-Nieto M., Santos-Rodriguez R., Kull M., Flach P. Classifier calibration: a survey on how to assess and improve predicted class probabilities//Machine Learning. -2023.-Vol.112. P.1-50. DOI 10.1007/s10994-023-06336-7.

16. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). - New York: ACM, 2016. - P.785-794. - DOI 10.1145/2939672.2939785.

17. Sahin E. K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest // SN Applied Sciences. - 2020. - Vol.2:1308. DOI 10.1007/s42452-020-3060-1.

18. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. CatBoost: Unbiased boosting

with categorical features // Advances in Neural Information Processing Systems (NeurIPS 2018). - 2018. - P.6638–6648. DOI 10.48550/arXiv.1706.09516

19. Shwartz-Ziv R., Armon A. Tabular data: Deep learning is not all you need // Information Fusion. - 2022. - Vol.81. - P.84-90. DOI 10.1016/j.inffus.2021.11.011.

20. Weber M., Domeniconi G., Chen J., Weidele D. K. I., Bellei C., Robinson T., Leiserson C. E. Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics // arXiv preprint. - 2019.- arXiv:1908.02591. DOI 10.48550/arXiv.1908.02591.

21. Johannessen F., Jullum M. Finding money launderers using heterogeneous graph neural networks // arXiv preprint. - 2023. - arXiv:2307.13499. DOI 10.48550/arXiv.2307.13499.

22. Hamilton W., Ying Z., Leskovec J. Inductive representation learning on large graphs//arXiv preprint. -2017.- P.1-19. DOI 10.48550/arXiv.1706.02216.

23. Probst P., Wright M. N., Boulesteix A.-L. Hyperparameters and tuning strategies for random forest // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. - 2019.- Vol.9(3): e1301. DOI 10.1002/widm.1301.

24. Nyberg O., Klami A. Reliably calibrated isotonic regression // Advances in Knowledge Discovery and Data Mining: PAKDD 2021. Lecture Notes in Computer Science. - Cham: Springer. -2021.- Vol.12712. - P.578–589. DOI 10.1007/978-3-030-75762-5_46.

25. Böken B. On the appropriateness of Platt scaling in classifier calibration // Information Systems. - 2021. - Vol. 95:101641. DOI 10.1016/j.is.2020.101641

Information about the authors

Azbergen T.K. - master's student, Kazakh-British Technical University, Almaty, Kazakhstan, e-mail: azbergen.torebek@mail.ru;

Tleubayeva A. O. - doctoral student, Astana IT University, Astana, Kazakhstan, e-mail: Arailym.tll@gmail.com;

Aituov A. T. - PhD, Kazakh-British Technical University, Almaty, Kazakhstan, e-mail: a.aituov@kbtu.kz;

Sabyrzhan Y. T. - master of Technical Sciences, Astana IT University, Astana, Kazakhstan, e-mail: sabyrzhan.yermakhan@astanait.edu.kz.

Сведения об авторах

Азберген Т. К. - магистрант, Казахстанско-Британский технический университет, Алматы, Казахстан, e-mail: azbergen.torebek@mail.ru;

Тлеубаева А. О.- докторант, Astana IT University, Астана, Казахстан, e-mail: Arailym.tll@gmail.com;

Айтуов А. Т. - PhD, Казахстанско-Британский технический университет, Алматы, Казахстан, e-mail: a.aituov@kbtu.kz;

Сабыржан Е. Т. - магистр технических наук, Astana IT University, Астана, Казахстан, e-mail: sabyrzhan.yermakhan@astanait.edu.kz.