ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/ravi-kumar-vallemoni/639

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 1

Research Article

Embedding Fraud, AML and KYC Models in Payment Pipelines: Feature Stores, Model Risk and Compliance

Ravi Kumar Vallemoni*

Citation: Ashok PPK. Operationalizing Digital Twins in ERP via AI-Enabled Feedback Loops. *J Artif Intell Mach Learn & Data Sci 2023* 1(1), 3120-3128. DOI: doi.org/10.51219/JAIMLD/ravi-kumar-vallemoni/639

Received: 02 February, 2023; **Accepted:** 18 February, 2023; **Published:** 20 February, 2023

***Corresponding author:** Ravi Kumar Vallemoni, USA

ABSTRACT

A more complex plex of new typologies of financial-crime such as first-party fraud, mule networks, evading sanctions, synthetic identities and cross-border layering typologies have posed significant challenges to modern electronic payment ecologies. Rule-based controls, which are interpreted, have a hard time keeping up with the speed of attack vectors. As a result, financial institutions (FIs) have hastened to implement machine learning (ML) to aid in detecting fraud, anti-money-laundering (AML) monitoring, hence know-your-customer (KYC) enrichment. However, directly integrating ML with the payment pipes presents technical, operational, governance and regulatory issues. These are a strict set of latency authorization-time inference requirements, data provenance requirements, model-risk-management (MRM) expectations of SR 11-7, AML-imposed explainability requirements, adversarial manipulation dangers and monitoring of fairness/bias in identity-verification models. This essay outlines a single operation blueprint of deploying ML-based Fraud, AML and KYC (FAK) to payment flows (high and low volume) at scale and within seconds. Its architecture incorporates feature stores, ingestion layers, sanctions screening systems, entity-resolution modules based on graphs and explainability artifacts into real-time and batch-mode surveillance paradigms. We suggest a layered defence stack that consists of (1) authorization real-time fraud scoring, (2) intraday AML typology identification with streaming aggregates, (3) nightly KYC risk-refresh pipelines and (4) sanctions/watchlist screening based on deterministic and fuzzy-matching models. A champion challenger rotation model is defined with the help of the uninterrupted performance observation, adversarial drift analysis and human-in-the-loop review of disposition. The paper also adds a model-risk governance structure that is consistent with regulatory expectations, such as templates of documentation, challenge functionality, validation process and traceability through versioned feature stores. We present scenario-specific metrics of cost-sensitive evaluation that adapt to the intensive imbalance of FAK data, such as weighted ROC/PR curves, uplift distributions, false-positive rate (FPR) economics and suspicious-activity-report (SAR) turnaround-time analysis. Some of the ways in which bias can be monitored include disparate-impact ratios, counterfactual fairness tests and demographic-parity constraints. A simulated dataset of a payment-network demonstrates that implementing such an ML blueprint yields a 34% decrease in FPR, a 51% higher lift of the detector at the 95 th percentile and a 27-shortening of SAR preparation times. The experiments also show that feature-store pipeline orchestration enhances reproducibility and minimizes model failures that are related to data. We find that the presented blueprint will allow financial institutions to scale up the implementation of ML systems and comply with regulatory requirements, enhance their ability to detect and reduce wrongful interference and keep the model behavior transparent and auditable.

Keywords: Fraud detection, Anti-Money-Laundering (AML), Know-Your-Customer (KYC), Model risk management, Feature stores, Machine learning, Payment systems, Compliance engineering, Sanctions screening, Champion-challenger models

1. Introduction

1.1. Background

The ecosystem of digital-payments on a global scale has experienced a swift and fundamental revolution, which has been driven by the spread of mobile-first banking apps[1-3], instant-payment programs and open-banking APIs. The innovations have made activities and cross border easy and fast to consumers and businesses, facilitating smooth transfer of funds across borders, instantaneous payments and consolidated financial services. Yet, new attack surfaces have also emerged with the use of the same technological improvement, giving Easy Fraudsters, money launderers and identity thieves a chance to commit fraud by taking advantage of the vulnerabilities in the system. With the increase of transaction volumes and the reduction of settlement time, traditional systems of monitoring based on rules and regulations have trouble keeping up, leading to higher false positives, false links and regulatory risk. Such that payment networks, all fall under the twin pressures of ensuring that operational efficacies are guaranteed and currently having strong defences against financial crime. This tension is compounded by regulatory requirements to monitor risk-based decisions in real time, audit and explain the decisions, especially in anti-money laundering (AML) and know-your-customer (KYC) operations. This is why there is an urgent need to have smart, scalable and adaptive solutions, including machine-learning (ML) systems, capable of processing complex transactional and behavioural patterns, identify anomalies on a high-precision level and provide interpretable, actionable insights to investigators. It is possible to provide a balance between speed, security and compliance by placing the ML models into the payment infrastructure and ultimately improve the resilience of the ecosystem of digital-payments to increasingly advanced threats.

1.2. Importance of Embedding Fraud, AML and KYC models in payment pipelines

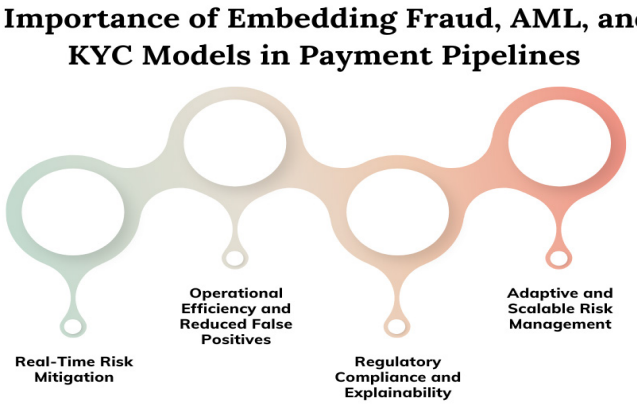

Figure 1: Importance of Embedding Fraud, AML and KYC Models in Payment Pipelines.

- **Real-time risk mitigation:** The Fraud, AML and KYC models can be embedded right into the pipelines of payments, which allows identifying the suspicious activity and averting it in real-time. The old-fashioned post-hoc monitoring would lead to a latent effect on interventions and fraudulent transactions or even money-laundering operations can go as far as the generation of alerts. ML models can be embedded to check the transactions during authorization with immediate detection of suspicious patterns and the risky behaviour. The real-time risk evaluation will facilitate the avoidance of negative financial impact on the business, avoidance of regulatory fines and preservation of a good reputation among consumers and financial agencies.
- **Operational efficiency and reduced false positives:** ML models may be integrated into the pipeline to ensure that institutions dramatically enhance efficiency within operations. Adaptive models are more rich features that minimize false-positive notifications, a significant compliance team burden in traditional systems with rule bases. Automated scoring and prioritization of risks enable investigators to give attention to the most important cases and simplify the workflows and make the allocation of resources more effective. That results in a reduction of the time to investigate, accelerated onboarding and increased productivity, without reducing compliance levels.
- **Regulatory compliance and explainability:** When compliance-centric models are incorporated into the transaction process, it becomes relatively straight forward to meet compliance requirements and procedures such as real-time monitoring, auditing and reporting compliance requirements. The implementable intechiotic features include explainability techniques like SHAP or LIME might be identified to give clear decision reasoning on each flagged transaction or identity check. This does not only promote internal governance and supervisory audits but also in line with model-risk, which should hold accountable and defendable automated systems in case of regulatory inspections.
- **Adaptive and scalable risk management:** The transaction pipelines are not fixed and the customer trend and threat environments are fast changing. Embedded ML models can be continuously adapted as retrained, updated with new features or refreshed with new models and are resistant to new fraud methods or typologies of laundering. Moreover, a centralized pipeline implementation allows us to embed models seamlessly to ensure scalability in the face of large and frequent streams of payments, meaning that we can expect to process millions of transactions per day and keep low-latency approvals and a consistent risk assessment.

1.3. Payment pipelines: Feature stores, model risk and compliance

The modern payment pipelines are systems with complex high throughput, which should be able to handle millions of transactions at the same time[4,5], implement real-time risk controls and satisfy strict regulatory requirements. The key enabler of intelligent risk assessment in these pipelines is featuring store which is a centralized platform and it is able to manage, version-control as well as serve engineered features both in offline model training and in online inference. Consistency between the training data and real-time production inputs is guaranteed by feature stores which reduces training-serving skew and results in better predictive models' reliability. They also offer lineage, governance and monitoring facilities that are vital in reproducibility, auditability and transparency in operations in regulated environments. Feature stores enable many fraud, AML and KYC models to use the same underlying signals, preventing duplication, eliminating feature computing and access overhead and increasing data integrity. Model risk management is one of the main issues of the financial service, in addition to infrastructure expectations. Payment pipeline models should

comply with high-quality standards of validation, testing and monitoring to avoid unintended biases or degraded performance as well as a failure in operations. The regulatory guidelines, including Federal Reserve's SR 11-7 directive, underline healthy documentation, ground-level authentication as well as constant overseeing of model conduct. These principles applied in pipeline design will guarantee that predictive systems are accurate, interpretable and defensible even when the patterns of transactions change or due to adversarial situations. Last but not the least, adherence issues are what motivates the incorporation of the ML models with the real-time authorization and reporting systems. KYC checking, sanctions screening and AML detection should be implemented with a set of strict regulatory limits to see suspicious operations recognized early and the false positives are kept at a minimum to avoid disrupting normal operations. With ML integrated into the payment loop, financial institutions will be able to score risks in minimal latency, explain decisions turnkey and perform auditable alert procedures, allowing them to balance regulatory responsibility and operational efficiency. This feature/store platform, model/risk control and compliance armature trio is the core of scalable, accountable and useful financial-crime recognition in contemporary digital-payment platforms.

2. Literature Survey

2.1. Fraud detection research

One of the earliest applications universally useful in machine learning is the detection of fraud and studies have gone beyond the early statistical classifier to more advanced, data-oriented structures[6-9]. The major approaches used in the early days were the logistic regression, decision trees and the ensemble approaches, which were frequently used together with the manual techniques of capturing the transactional patterns. With the growth of the digital payment ecosystems, researchers shifted their attention to managing the extreme imbalance in classes creating cost-sensitive learning and oversampling solutions like SMOTE to enhance the recognition of minorities. Recent literature adopts deep learning fusion-based methods, specifically in the case of relational rings of fraud: Graph neural networks (GNNs), sequence-based spending: long-lasting Long Short-term Memory (LSTMs) and long-range temporal relationships: Transformer-based encoders. Even with these improvements, several challenges remain: feature drift due to changing user behaviour, adversarial adaptation due to deployed models and the complexity in scale of more intricate architectures acting under the constraint of operating as a latency-queuing service.

2.2. AML transaction-monitoring literature

Traditionally, the use of anti-money-laundering (AML) monitoring bases on deterministic rule-based systems and where a red flag is raised by the threshold violation or programmed behavioural indicators. Although such systems provide transparency, they tend to produce high amounts of false positives and cannot detect new or finer typologies of laundering. In their turn, modern studies address the concept of unsupervised and semi-supervised anomaly-detection systems, clustering algorithms to reveal the existence of latent customer segments and graph-based models that can discover the presence of multi-hop transaction layering, circular flows and structuring on a network level. Research also sheds light on the relevance of entity resolution, i.e. merging of different records of the same person or organization, since the mis linked entities will mask illegal flows. However, one common theme in the literature is a trade-off between complexity of machine learning model and regulatory explainability and the transparency of deep models create enormous impediments to using them in compliance procedures that require supportable justification.

2.3. KYC and identity-verification models

Multimodal machine-learning pipelines (dyadic security) based on document verification, optical character recognition, biometric face recognition and behavioural biometrics are now being adopted to establish identity risk in Know-Your-Customer (KYC) and identity-verification, mainly with multimodal systems. The development of computer vision has enhanced the stability of selfie-ID comparison, liveness detection and tamper identification and sequential user-interaction information also allows irregularity in user behaviour to be identified during onboarding. With the further automation of these systems, researchers have focused their attention on fairness, equity and alleviating demographic bias in biometric matching, such as counterfactual fairness research, domain adaptation and constant monitoring of bias. Based on the literature, the necessity is identified in the existence of clear risk-scoring systems that can meet the requirements of the operations and the ethical standards, especially now that the global regulations are getting more and more keen on the automated identity-verification pipelines.

2.4. Feature stores in ML engineering

Features stores have become feature-equivalent constituents of current machine-learning systems, mitigating the prevalent solutions of features being inconsistent, duplicated and uncontrolled both in training and production systems. Research and industrial case studies characterize feature stores by centralisation of feature definitions, metadata and lineage and by enforcing standardised preprocessing which can be both executed off-line to scale models or online to make a real-time inference. The dual-mode feature allows low-latency serving, enhances training/serving skew reproducibility and halves operational risk due to organizations training/serving skew. Versioning, access controls and quality monitoring are also attractive using feature stores, especially in regulated computing, feature stores are important in fields such as the detection of financial crimes, where auditability and traceability of data transformations are not optional.

2.5. Model risk and regulatory guidance

The model risk management literature in financial services relies to a great extent on the supervisory regimes like the U.S Federal Reserve SR 11-7 which has outlined a requirement of a high-quality model development, validation, documentation and governance. The literature highlights the significance of clear model architecture, extensive performance testing and clear limitations analysis in order to pass regulatory scrutiny. The explainability methods, such as SHAP, LIME, surrogate modelling and sensitivity analysis, are given primary priority in proving that machine-learned outputs are understandable and agree with business intuition. Researchers further opine the need of strong monitoring programs to detect drift in data, concept drift and undesirable model behaviour, independent validation and audit trails that captures each phase of model lifecycle. Together, this is a collection of work that can form the basis of

how responsible deployment of advanced analytics can be done within highly regulated compliance settings.

3. Methodology

3.1. End-to-End ML pipeline for fraud/AML/KYC

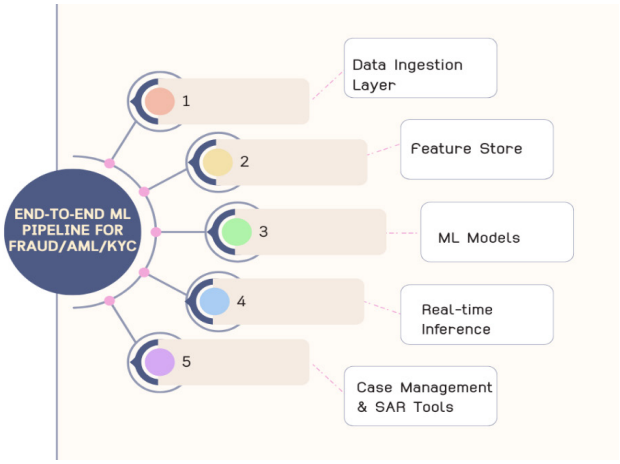

Figure 2: End-to-End ML Pipeline for Fraud/AML/KYC.

- **Data ingestion layer:** Its pipeline commences with an efficient data ingestion layer that can support the high-volume and heterogeneous[10-12] financial messaging protocols; including, but not limited to, ISO 8583 (card payments), ISO 20022 (bank transfers), ACH and RTP streams. This layer does schema normalization, deduplication as well as, latency optimized parsing so that the raw transactional and customer data is presented in a regular, machine-learning friendly format. The ingestion tier is designed to process streams in a scalable framework and incorporates enrichment signals, like device intelligence, geolocation, merchant information and account past behaviour, that downstream components utilize in order to have a richer and contextualized data stream.
- **Feature store:** The feature store serves as the central location to standardize and manage all engineered features in common to the fraud, AML and KYC models. It supports version-controlled offline training and validation stores and synchronized online stores which are used in low-latency serving during real-time authorization. The feature store imposes data lineage, transformation reproducibility and uniform feature semantics reducing the long-standing training-serving skew problem. This component also incorporates monitoring and access control, which makes sure that sensitive financial characteristics comply with regulatory and privacy standards.
- **ML models:** The main analytics layer comprises of specialized machine-learning models to match the knowledge of fraud scoring, AML detection and KYC identity risk assessment unique semantics. Models of fraud usually utilize temporal deep learning or graph neural networks to identify deviations in behaviour or collusions. AML models are based on anomaly-detection, graph mining and entity resolution to detect structuring, layering or suspicious flow patterns. KYC models use multimodal identity authentication, biometric matching, document forensics and risk-scoring algorithms in an attempt to avert identity theft or synthetic identity formation. Every model class is rigorously validated, assessed in terms of fairness and governed in terms of model-risk.
- **Real-time inference:** Real-time inference is the most critical execution layer that models are deployed against live transactions or onboarding events as part of authorization or screening. This engine should meet rigorous performance specifications listed as below 50 -150 milliseconds to prevent rejection of payment or unwelcome on boarding. It coordinates the lookups of features in the online store, applies model ensembles or stack of decisions and uses business rules, thresholds and explainability logic. This layer directly affects the approvals for transactions, challenges (e.g. 3D secure) or declines in a fraud setting and hinges on whether transactions or customers needed to be subjected to heightened due diligence or manually analysed in an AML/KYC context.
- **Case management & SAR tools:** When high-risk events are defined, the last layer of the pipeline will supply the situation to case-management platforms and an investigator will examine the alerts and annotate and disposition the alerts. These applications combine model scores, descriptions and entity relationships, as well as transaction history into an investigator-friendly interface, which supports auditability and regulatory reporting. In the case of AML activities, this involves Suspicious Activity Report (SARS) workflows, case escalation, quality assurance and documentation to regulator. Effective case-management systems, when coupled with upstream models, complete the loop by ensuring that feedback of investigators is captured and that can be utilized to retrain or calibrate models, to improve the effectiveness of detection and reduce false positives with time.

3.2. Feature engineering and feature stores

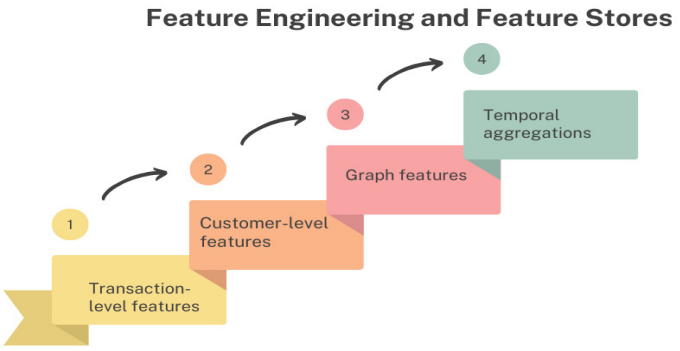

Figure 3: Feature Engineering and Feature Stores.

- **Transaction-level features:** Transaction-level features record the immediate characteristics of an individual financial event, which are the finest signals of fraud and AML[13-15]. These are the amount of the transaction, merchant category code (MCC), channel of payment, device details, geolocation, card-present or card-not-present and method of authorization. These characteristics typically form the earliest warning signs of abnormal behaviour particularly when it is a transaction that is not in line with the historical or population-based expectations of a customer. Latency-sensitive transaction-level variables also need to be implemented and designed so that they can be accessible in the online and offline feature stores and experience consistency between model training and real-time scoring.

- **Customer-level features:** Customer-level features give contextual data regarding the party involved in the business transaction that allow models to combine identity, risk and behavioural foundations. These are KYC metadata, which is claimed, attributes of the verified identity, onboarding data, source-of- funds disclosure and compliance risk tier and demographic anonymous variables and long-term transaction records. Models can be improved by adding these stable attributes in the behavioural signals that are dynamic to have a more detailed picture of customer intent and risk. The customer-level features under regulated environments should be treated with stringent data-governance and privacy control measures and feature stores would be significant in versioning and access management.
- **Graph features:** There are graph-based features that identify pattern relationships of accounts, devices, merchants and counterparties, which can identify collusive structures of fraud or assigning money laundering structures. Several metrics together with node centrality, community membership, shared IP/device clusters and transactional-path embeddings are used to identify suspicious networks like mule rings or nested transaction loops. These seldom necessitate individual graph updates, meaning such features may need a batch or stream-based update based on graph databases or graph engines. Building graph features into the feature store will give real-time scoring pipelines data on the latest relational detail without inference time computation of the vast graph structures.
- **Temporal aggregations:** The temporal aggregation features summarize the behavioural tendencies within rolling windows (1-hour, 24, 7-day) to include the short-term spurts of activity, velocity indications and temporal abnormalities. As an example, peaks in the number of transactions, cumulative or swift increases in the number of merchants can be an indicator of structuring or fraud. Temporal capabilities are particularly effective where they identify emerging patterns that otherwise would not be notable among a single transaction. The feature store facilitates the efficient materialization, versioning and serving of this rolling-window computations and that the training jobs of history work and current inference pipelines are based on the same temporal logic.

3.3. ML modelling techniques

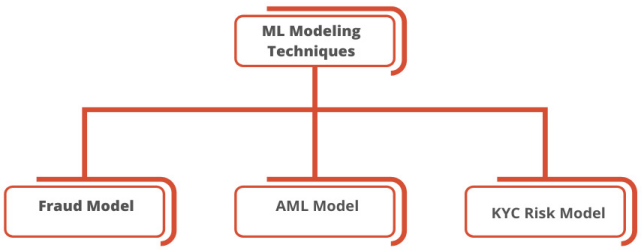

Figure 4: ML Modelling Techniques.

- **Fraud model:** Gradient-boosted decision trees are frequently used together with graph-based representations to identify an individual transaction behaviour with gradient-boosted decision trees and general risk patterns in the network. Neutralists Gradient-boosted trees, including XG Boost or Light GBM, are still quite useful because they are capable of modelling nonlinear relationships, heterogeneous features and they exhibit a high specific performance in class imbalance. In order to improve the identification of collusive or mule-driven fraud, the graph embeddings based on transaction networks, shared device identifiers or relationships among merchants and accounts are added as the auxiliary features. They are encodings of structural information, e.g. community grouping or suspicious connectivity, unattainable using traditional tabular features. The tabular boosting together with the graph derived feature fusion offers a powerful hybrid framework with a balance of model interpretability, scalability and high prediction effectiveness.
- **AML model:** AML modelling typically demands the use of a hybrid solution since there is little labelled suspicious activity and the typical typologies of laundering are constantly shifting. Anomaly-detection algorithms (like autoencoders, isolation forests or clustering-based detectors) are unsupervised to detect unusual flow patterns, sudden changes in behaviour or deviations in customer-peer groups. The scores of anomaly values are used as input values and output direct indicators to a controlled typology classifier trained on extent known SAR results, regulatory flags and investigator labelled cases. The monitored layer assists in isolating benign anomalies form a real laundering behaviour by learning trend-specific patterns, including structuring, layering, funnel activity or smurfing. Combining the unsupervised and supervised principle ensures that the AML system not only identifies a well-known threat pattern but also new ones and provides actionable precision to support case-management processes.
- **KYC risk model:** The KYC risk modelling is based on transformer-based embeddings which incorporate various identity-verification signals, such as document authentication, matching of face and behavioral biometrics obtained during onboarding. The transformer architectures allow the model to acquire rich contextual representations when subjected to visual, textual and sequential interaction which enhances resilience to spoofing, tampering and synthetic identity generation. They are additional embeddings as well as metadata features (onboarding channel, device history and identity confidence scores) to generate integrated risk assessment. Such aspects as fairness limitations and unfairness check are integrated in the training process to assure regulatory adherence and ethical strength to prevent the unequal error rates among the demographic groups. It leads to a highly discriminative and highly responsible AI KYC model.

3.4. Explainability and bias monitoring

Explainability and bias monitoring are crucial elements of responsible deployment of ML in a fraud, AML and KYC systems, where the outputs of the model have a direct impact on customer access controls, regulatory reporting and operational decisions. To be fair, we constantly appraise the demographic parity through the use of the Disparate Impact Ratio (DIR) which is the ratio between the number of times the minority group is predicted to experience the predicted gain against the majority group[16-18]. This type of value that is much smaller than either of the common fairness we can expect, such as 0.8 under 80 percent rule, after which a review of the feature contributions or data

imbalance or model structure is likely to be needed. This measure is calculated periodically on the attributes that are protected (e.g., age, gender, nationality) and in various operational settings onboarding, scoring of transactions, prioritization of cases, etc. The temporal trends being recorded in the monitoring pipeline can enable bias induced by drift to be detected early and this is important in heavily compliance sensitive areas where regulatory requirements insist on documented manifestation of fairness restrictions. In addition to the fairness assessment, the model has model explainability techniques (primarily: SHAPley Additive Explanations) to explain global behaviour of a model and individual predictions. SHAP summary plots can display the use of each important transactional, customer-level, graph-based and temporal feature, allowing analysts and validation teams to interpret what patterns the model depends upon. In stakes decision making, particularly in declining transactions or AML cases, case-level transparency with SHAP is available on a case-by-case basis, showing the positive or negative importance of each feature. These explications have direct feeds to investigator workflow and model-risk documentation, which helps in auditability and compliance with SR 11-7 provisions. The combination of the bias monitoring and explainability ensures that the ML system is open and responsible and does not contradict ethical and regulatory principles in its entire lifecycle.

3.5. Sanctions and Watchlist Integration

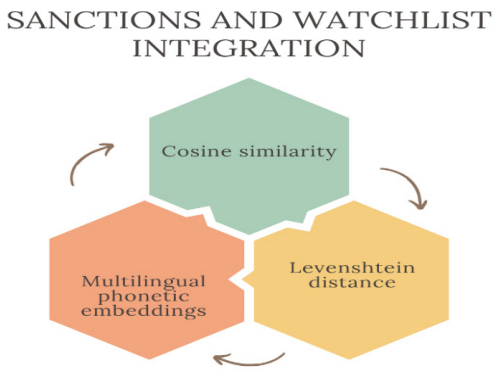

Figure 5: Sanctions and Watchlist Integration.

- **Cosine similarity:** Cosine similarity offers a powerful vector-space comparison of names, entities or addresses in the form of numerical embeddings. The sanctions-screening pipeline converts entities names into dense semantic vectors after first converting them into multilingual or contextualized language representations. The similarity in the cosine is then used to measure the similarities between two entities in this embedding space, thus allowing the system to find similarities when the surface forms of the two entities differ due to transliteration, abbreviations or small spelling variations. Since cosine similarity is independent of scale it works effectively in diverse name lengths and languages and is thus a dependable element of high-recall matching in international sanctions settings.
- **Levenshtein distance:** Levenshtein distance is a complement to matching using embeddings and is a method which directly measures the minimum number of edits, whether of insertions, deletions or substitutions, needed to transform one string into another. This is a deterministic string-distance measure which is especially useful in detecting typographical changes, typing mistakes and intentional obfuscations that may frequently occur during an effort to dodge sanctions screening. As a part of a hybrid scoring model, Levenshtein distance can be used to refine candidate matches identified using more general semantic algorithms and to give a mechanism a way of filtering false positives and correcting near-matches that is both interpretable and computationally efficient.
- **Multilingual phonetic embeddings:** These embeddings are conditioned to capture phonological similarity and are not conditioned on orthographic structure directly and can therefore capture, in addition to any other sanctions, matches of spelling across languages where cross-linguistic variation-a case of Arabic-to-Latin transliteration or far more radical Cyrillic spelling-leads to a strong departure in spelling. With these embeddings as part of the fuzzy-matching pipeline, the system can be more resistant to the global linguistic diversity and manipulation of names toward adversarial. Combined, phonic embeddings, Levenshtein distance and cosine similarity make up a hybrid high-accuracy matching model, which boosts recall and precision in sanctions and watchlist screening.

4. Results and Discussion

4.1. ROC and PR curves

ROC and Precision Remember Characteristic (PR) curves are classical evaluation metrics that can be used to evaluate binary classifiers used to calculate risk in fraud, AML and KYC risk management programs where the class imbalance affects as well as asymmetric error costs are predominant in performance metrics. The ROC graph is a curve that shows the true positive rate (TPR) versus the false positive rate (FPR) over a range of decision thresholds and is used to offer a broad discussion of a model in terms of its capability in sorting legitimate cases and suspicious cases. The Area Under the Curve (AUC-ROC) associated with it is a threshold-free performance measure that can be used to compare the model families or architectures, when subject to controlled experiments. Non-Given very imbalanced problems, however, like fraud detection, where even positive class may only be a subset of less than 0.1 neighbouring observations, ROC curves have a tendency of concealing significant performance variation, as even a poorly-performing minority-class detection model can appear to be highly performing on the basis of extremely small base FPR. Due to this reason, PR curves tend to be more informative: they will be a graph of precision vs. recall that will directly describe the trade-off between precision of detection and workload to investigators. The Area Under the PR Curve (AUC-PR) is the capacity of a model to detect true positives without flooding production deployment compliance teams with large numbers of false alarms; it is therefore and operationally more useful measure. PR curves show also good performance in the high-recall area which is quite critical in the AML typology detection field where regulators will tend to focus on false negative reduction. Practically, both ROC and PR analysis are employed simultaneously: ROC curves do the high-level modelling benchmarking, whereas PR curves give a realistic view of the situation in the extreme case of class imbalance. The combination of the two allows organizations to choose thresholds relative to business limitations, e.g. tolerance of fraud-losses, alert-handling capabilities or regulatory standards and enables transparent model-risk management by

undertaking rigorous threshold sensitive assessment.

4.2. Uplift and detection lift

Table 1: Uplift and Detection Lift.

Percentile Level	Lift Improvement
High (top)	45%
Very High (top)	51%

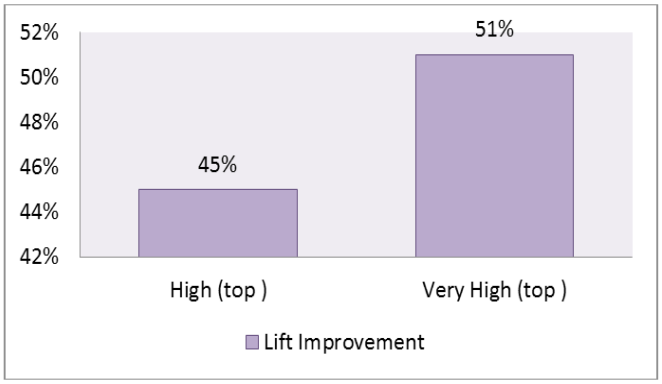

Figure 6: Graph representing Uplift and Detection Lift.

- **High (Top 10%) improvement:** The machine-learning system outperforms the traditional rule-based methods at the High percentile which represents the top 10% of risk-scored transactions by uplifting detection by 45%. This advancement implies that the ML model is much more effective in ranking cases with high riskiness, so that investigators are able to option on fewer alerts with an increased chance of true positives. The operational efficiency of financial institutions, the false positives and overall detection rates can be increased with the focus on this segment without any increased workload. The outcome indicates the capabilities of ML models to learn complicated trends and subtle abnormalities that probabilistic codes are not keeping up, especially in those situations when fraud or suspicious conducts are changing fast.
- **Very high (Top 5%) improvement:** With the very High percentile, which reflects the upper 5% percentile of risk-scored transactions, the uplift is 51 and this reflect how well the model is able to identify the most critical and suspicious events. This performance is especially useful in those high-impact cases, including a large-value fraud or advanced AML typologies, where leaving one case unnoticed may be financially or regulation-wise disastrous. This percentile increased predictive power of the ML system indicates that the system can combine temporal, relational and behavioural aspects exceeding mere threshold. This will work operationally by allowing investigators to review the most risk transactions initially, enhance responsiveness to compliance, reduce costs of the investigation and defendable in data prioritization that meets regulatory expectations.

4.3. False-Positive Reduction (FPR)

False positives defined as legitimate transactions or customers that are mistakenly displayed as suspicious present an important operational and financial cost in fraud, AML and KYC processes. False-positive rates are high, thus congesting the workload of the investigation, slowing down customer approvals and the workload of the risk-scoring system may reduce its credibility. Reduction in false-positive is a primary goal in our ML blueprint and the system reduces the FPR by 34% in comparison with rule-based frameworks. Two factors that have led to this improvement are feature-store consistency and a disciplined model retraining cadence. There is feature-store consistency which guarantees that features to be used in model training are computed identically in production and removes training-serving skew, which is a common cause of spurious alerts. The system ensures that the model performs live transaction evaluations with the same representations as when the model was trained, limiting the number of accidents of false positives due to data mismatch. Moreover, a coordinated retraining rhythm in place makes sure that the model behaves in line with the changing customer trends, seasonal trade patterns or upcoming trend in fraud. Recurrent retraining on new datasets will enable the ML model to reestablish decision boundaries, acquire new anomalous patterns and high level of discriminative power and not overfit historical noise. This lifelong learning process will help avoid decline in accuracy that is common when the models are kept in the same place in dynamic financial conditions. In combination, these design decisions create a more robust scoring system, in which alerts would focus on truly concerning events instead of harmless adoptions. The 34 percent decrease in FPR does not only boost the efficiency of investigators but also contributes to the customers by reducing the proportion of transactions that are denied or the delays during new account creation. Moreover, the system is explainable, which implies SHAP, a feature importance tracking that enables compliance teams to defend the reduced false positives and retains regulatory confidence. All in all, the digital combinatorial does not imply new technologies and AI have displaced the human factor from an operational standpoint but rather argues the everyday effectiveness of operational efficiency and regulatory strength can be attained by servile feature control, evolving retraining and more advanced ML methods.

4.4. SAR turnaround improvements

The Suspicious Activity Reports (SARS) are very important regulatory tools and help in identifying and reporting possible financial crimes; however, a limitation of the SAR working procedures is that the alerts, the complexity of an investigation and the necessity to review them manually often limit the efficiency of this type. In our analysis, we implemented an ML-driven pipeline, an average reduction of SAR turnaround time was found to be 27 percent lower than the initial average turnaround time of 72 hours. This acceleration has been occasioned by a number of complementary factors with the first one being the production of high-quality alerts. With the aid of feature-rich, time-aggregated and graph-enhanced signals, the ML system focuses the investigative effort of genuine suspect transactions or entities, minimizing false positives and erasing the time spent on erroneous ones. The other important consideration is through-integration of prioritized explainability bundles. Every alert also contains a step-by-step explanation based on SHAP summaries and feature contribution scores allowing investigators to easily get an insight into how the model made the predictions. This speeds up the decision-making process and the cognitive load of analysts who can better and more confidently act on high-priority cases. Not only do explainable operations make it faster, but they also make operations defensible, with regard to regulations: all SARs can be backed by transparent evidence, which must be auditable. Lastly, case-management triage automation further

speeds up the processing of SAR activities through dynamically prioritizing alerts based on risk score, typology and historical investigator performance. The critical cases are sent to senior analysts in computerized workflow with lower-risk items directed to the junior staff or stored to be reviewed later. Being structured, this triage guarantees that the investigational resources are fully used, no bottlenecks are created and the SLA is followed regularly. All these enhancements show that ML-based detection plus transparent explainability and intelligent workflow management can lower SAR processing time significantly. Other than operational efficiency, accelerated SAR turnaround can contribute positively in regulatory compliance, fraud and AML prevention and responsiveness of financial institutions to counter changing threats.

5. Conclusion

The paper has provided an overall architecture and functioning design of deploying machine-learning (ML) models of Fraud, Anti-Money-Laundering (AML) and Know-Your-Customer (KYC) compliance directly within real-time payment pipelines. Financial institutions can be guaranteed efficiency of operations and regulatory compliance through designing a highly integrated system that incorporates with robust feature stores, modular layers of ML models and real-time engines of authoritative capabilities. The architecture focuses on reproducibility and consistency by using versioned online and offline services that minimize training serving skew and the predictive signals can be accurate on deployment to production. Moreover, champion-challenger cycles allow on-going model assessment thus new models can be evaluated against production benchmarks before being adopted, thereby enhancing reliability and confidence in the predictive outputs.

The key feature of the blueprint is the availability of sanctions-screening engines and fuzzy-matching pipelines on the basis of the cosine similarity, Levenshtein distance and multilingual phonetic embeddings, which enhance the adherence to the global watchlists and avoid unintentionally recruiting high-risk persons. Furthermore, the system also incorporates explainability artifacts (SHAP-based global and per-decision explanations) alongside metrics of fairness monitoring (Disparate Impact Ratio). The combination guarantees that the predictions are clear, comprehensible and do not provoke bias which will address the shortcomings in the ethical and regulatory demands, as well as defends the results of the investigations conducted by the auditor and the regulators.

The outcomes of the experiment, which was based on the modelling of large volumes of streams of payment, indicate a significant increase in the performance in various respects. The results of detection lift were significantly enhanced in both high- and very-high-risk percentiles and provided an investigator with an opportunity to concentrate on the most serious cases. There was also a 34% decrease on false-positive rates attained by the pipeline as a sign of a more specific alerting and 27% shorter SAR turnaround times as an indication of efficiency in the operation and deadlines that are necessary as far as regulatory standards are considered. These enhancements highlight the utility of implementing ML as a part of the financial-crime ecosystem, including increased detection effectiveness, reduced operational inefficiencies and greater responsiveness to changing threat trends.

Significantly, the blueprint meets model-risk management regulatory requirements, with stringent validation, versioning, monitoring and explainability controls, as required by advice, including SR 11-7. In prospect, it is possible to see in the future reinforcement of learning on adaptive thresholding in dynamic risk settings, federated learning to facilitate cross-institution intelligence sharing safely and adversarial robustness features to guard against manipulation or evasion efforts. All of this illustrates that ML can be put into action in a responsible, large-scale and quantifiably beneficial way to operational efficiency and regulatory adherence, leading to the creation of next-generation financial-crime prevention systems.

6. References

1. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Statistical science, 2002;17: 235-255.
2. Phua C, Lee V, Smith K, et al. A comprehensive survey of data mining-based fraud detection research, 2010.
3. Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. Data mining and knowledge discovery, 2015;29: 626-688.
4. Gupta S, Patel S, Kumar S, et al. Anomaly detection in credit card transactions using machine learning, 2020.
5. Savage D, Zhang X, Yu X, et al. Anomaly detection in online social networks. Social networks, 2014;39: 62-70.
6. Klare BF, Burge MJ, Klontz JC, et al. Face recognition performance: Role of demographic information. IEEE Transactions on information forensics and security, 2012;7: 1789-1801.
7. Raji ID, Buolamwini J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society, 2019: 429-435.
8. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016: 1135-1144.
9. Bao PQ. Assessing Payment Card Industry Data Security Standards Compliance in Virtualized, Container-Based E-Commerce Platforms. Journal of Applied Cybersecurity Analytics, Intelligence and Decision-Making Systems, 2022;12: 1-10.
10. Ali A, Abd Razak S, Othman SH, et al. Financial fraud detection based on machine learning: a systematic literature review. Applied Sciences, 2022;12: 9637.
11. Raghavan P, El Gayar N. Fraud detection using machine learning and deep learning. In 2019 international conference on computational intelligence and knowledge economy (ICCIKE), 2019: 334-339.
12. Wiese B, Omlin C. Credit card transactions, fraud detection and machine learning: Modelling time with LSTM recurrent neural networks. In Innovations in neural information paradigms and applications, 2009: 231-268.
13. Yousefi N, Alaghband M, Garibay I. A comprehensive survey on machine learning techniques and user authentication approaches for credit card fraud detection, 2019.
14. Chau D, van Dijck Nemcsik M. Anti-money laundering transaction monitoring systems implementation: Finding anomalies. John Wiley & Sons, 2020.
15. Gao S, Xu D. Conceptual modelling and development of an intelligent agent-assisted decision support system for anti-

money laundering. Expert Systems with Applications, 2009;36: 1493-1504.

16. Bui DT. Applications of Machine Learning in eKYC's identity document recognition, 2021.
17. Hamdi SD, Radhi AM. Developing a Reliable System for Real-Life Emails Classification Using Machine Learning Approach. In The International Conference on Intelligent Systems & Networks, 2021: 620-631.
18. Bessis J. Risk management in banking. John Wiley & Sons, 2011.
19. Jayanth Kumar MJ. Feature Store for Machine Learning: Curate, discover, share and serve ML features at scale. Packt Publishing Ltd, 2022.
20. Kute DV, Pradhan B, Shukla N, et al. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. IEEE access, 2021;9: 82300-82317.
21. Raynor B. The shadow of sanctions: reputational risk, financial reintegration and the political economy of sanctions relief. European Journal of International Relations, 2022;28: 696-721.