

Data Science Canvas		Project:	Advanced Machine Learning for Nigerian Banking Fraud Detection using NIBSS Dataset					
		Team:	Data Conduits					
Problem Statement				Execution & Evaluation		Data Collection & Preparation		
<b>Business Case &amp; Value Added</b> <p>Nigerian banks lost ₦17.67B (~\$216M) in 2023, with fraud cases rising 23% YoY. Traditional rule-based systems struggle with evolving fraud patterns and generate high false positives. This project delivers a machine-learning-driven fraud detection system that:</p> <ul style="list-style-type: none"><li>i. Reduces fraud losses through early detection</li><li>ii. Minimizes false positives</li><li>iii. Automates low-risk decisions to reduce manual review workload</li><li>iv. Enables fraud teams to focus on high-risk, high-value cases</li><li>v. Improves compliance and customer trust with explainable predictions</li></ul>	<b>Model Selection</b> <p>Given the extreme class imbalance (0.3% fraud cases), the solution uses:</p> <ul style="list-style-type: none"><li>Gradient boosting models (LightGBM, XGBoost) for non-linear patterns</li><li>Ensemble stacking to leverage complementary feature importance</li><li>LSTM embeddings to capture short-term temporal transaction patterns</li><li>Optuna for hyperparameter and ensemble weight optimization</li><li>Precision-Recall threshold tuning to maintain extremely low FPR</li></ul>	<b>Model Requirements</b> <p>Performance: <math>F1 \geq 0.90</math> (achieved 0.8847), <math>AUC-ROC \geq 0.95</math>, <math>FPR &lt; 0.1\%</math> (achieved 0.001%)</p> <p>Explainability: SHAP-based interpretability for audit and compliance</p>	<b>Skills</b> <p>Data Engineering: Pipeline design for high-volume transaction data</p> <p>Data Science: Feature engineering (temporal, velocity, interaction), ensemble modeling</p> <p>ML Engineering: Optuna tuning, model serving, real-time scoring pipelines</p> <p>Risk/Fraud Expertise: Interpret patterns, label verification, regulatory compliance</p>	<b>Model Evaluation</b> <p>Performance is assessed using <math>F1</math>, <math>AUC-ROC</math>, <math>PR</math> <math>AUC</math>, and precision-recall balance due to extreme imbalance. The ensemble achieved:</p> <p>AUC: 0.9638 <math>F1</math>: 0.8847</p> <p>FPR: 0.001% (100x better than requirement)</p> <p>Confusion matrix analysis guides threshold adjustment. Real-time monitoring includes drift detection, data quality checks, and alerting when fraud distribution shifts.</p>	<b>Data Storytelling</b> <p>Target Group Reqs Clear, business-focused insights on fraud patterns affecting Nigerian payment channels.</p> <p>Operationally relevant metrics (fraud hotspots, transaction patterns, risk levels by customer/merchant segment).</p> <p>Actionable recommendations that regulators, banks, and fintech operators can implement immediately.</p> <p>Simple, interpretable visuals that work for mixed stakeholders (analysts, compliance teams, executives).</p> <p>Traceability and transparency in how fraud was detected, especially for regulatory review.</p>	<b>Data Selection &amp; Cleansing</b> <p>Relevant features like: Temporal: hour, day, rolling time gaps Velocity: <math>tx\_count\_24h</math>, <math>amount\_sum\_24h</math> Behavioral: <math>ratio\_to\_customer\_max</math>, deviation features Interaction: composite risk scores</p> <p>Data cleansing included: Handling missing timestamps Outlier detection for extreme transaction amounts Encoding categorical fields (channel, location, merchant)</p>	<b>Data Collection</b> <p>Real-time data from mobile banking, transfers, POS, ATM, web</p> <p>Fraud labels from investigation teams</p> <p>Need for sequence data per customer (for LSTM/transformer models)</p>	
<b>Data Landscape</b> <p>NIBSS Fraud Dataset (Kaggle) — synthetic data that is representative of transaction data reflecting Nigerian banking characteristics.</p> <p>Synthetic data is chosen as banking data is sensitive, and personal.</p>		<b>Software &amp; Libraries</b> <p>Python pandas, NumPy scikit-learn LightGBM, XGBoost, CatBoost Optuna PyTorch/Keras for LSTM SHAP Matplotlib/Seaborn</p>			<b>Data Integration</b> <p>Lead with key fraud insights (where, how, and why fraud occurs in the Nigerian context). Use clean visuals (heatmaps, trend lines, risk scores) instead of technical plots. Translate findings into operational actions: improved KYC checks, velocity limits, merchant risk tiers, model cutoffs. Highlight local relevance: mobile money patterns, high-risk time windows, account takeover behaviors typical in Nigeria.</p>		<b>Explorative Data Analysis</b> <p>Fraud prevalence: 0.3% Fraud concentrated 1 AM – 4 AM High-risk channels: mobile, web Fraud values skew higher than legitimate ones Lagos &amp; Abuja show highest fraud density Temporal and velocity patterns drive strong predictive power</p>	