# Data Science Canvas

**Project:** Nigerian Banking Fraud Detection System

**Team:** Data Conduits

## Problem Statement

**Business Case & Value Added**
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:
i. Reduces fraud losses through early detection
ii. Minimizes false positives
iii. Automates low-risk decisions to reduce manual review workload
iv. Enables fraud teams to focus on high-risk, high-value cases
v. Improves compliance and customer trust with explainable predictions

**Data Landscape**
NIBSS Fraud Dataset (Kaggle) — synthetic data that is representative of transaction data reflecting Nigerian banking characteristics.
Synthetic data is chosen as banking data is sensitive, and personal.

**Model Selection**
Given the extreme class imbalance (0.3% fraud cases), the solution uses:

Gradient boosting models (LightGBM, XGBoost) for non-linear patterns

Ensemble stacking to leverage complementary feature importance

LSTM embeddings to capture short-term temporal transaction patterns

Optuna for hyperparameter and ensemble weight optimization

Precision–Recall threshold tuning to maintain extremely low FPR

**Model Requirements**
Performance: F1 ≥ 0.90 (achieved 0.8847), AUC-ROC ≥ 0.95, FPR < 0.1% (achieved 0.001%)

Explainability: SHAP-based interpretability for audit and compliance

**Software & Libraries**
Python
pandas, NumPy
scikit-learn
LightGBM, XGBoost, CatBoost
Optuna
PyTorch/Keras for LSTM
SHAP
Matplotlib/Seaborn

**Skills**
Data Engineering: Pipeline design for high-volume transaction data

Data Science: Feature engineering (temporal, velocity, interaction), ensemble modeling

ML Engineering: Optuna tuning, model serving, real-time scoring pipelines

Risk/Fraud Expertise: Interpret patterns, label verification, regulatory compliance

## Execution & Evaluation

**Model Evaluation**
Performance is assessed using F1, AUC-ROC, PR AUC, and precision–recall balance due to extreme imbalance. The ensemble achieved:

AUC: 0.9638
F1: 0.8847

FPR: 0.001% (100× better than requirement)

Confusion matrix analysis guides threshold adjustment.
Real-time monitoring includes drift detection, data quality checks, and alerting when fraud distribution shifts.

**Data Storytelling**
Target Group Reqs
Clear, business-focused insights on fraud patterns affecting Nigerian payment channels.
Operationally relevant metrics (fraud hotspots, transaction patterns, risk levels by customer/merchant segment).
Actionable recommendations that regulators, banks, and fintech operators can implement immediately.
Simple, interpretable visuals that work for mixed stakeholders (analysts, compliance teams, executives).
Traceability and transparency in how fraud was detected, especially for regulatory review.

How?
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.

## Data Collection & Preparation

**Data Selection & Cleansing**
Relevant features like:
Temporal: hour, day, rolling time gaps
Velocity: tx_count_24h, amount_sum_24h
Behavioral: ratio_to_customer_max, deviation features
Interaction: composite risk scores

Data cleansing included:
Handling missing timestamps
Outlier detection for extreme transaction amounts
Encoding categorical fields (channel, location, merchant)

**Data Integration**
Different data sources (transaction logs, customer profiles, merchant metadata, fraud investigation labels) are unified into a centralized feature store.

**Data Collection**
Real-time data from mobile banking, transfers, POS, ATM, web

Fraud labels from investigation teams

Need for sequence data per customer (for LSTM/transformer models)

**Explorative Data Analysis**
Fraud prevalence: 0.3%
Fraud concentrated 1 AM – 4 AM
High-risk channels: mobile, web
Fraud values skew higher than legitimate ones
Lagos & Abuja show highest fraud density
Temporal and velocity patterns drive strong predictive power