INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

DA 204o: Data Science in Practice
Course Project Proposal
Team: Data Conduits
Advanced Machine Learning for Nigerian Banking Fraud Detection using NIBSS Dataset

Novoneel Chakraborty, CDPG, FSID, cnovoneel@iisc.ac.in
Sarthak Sharma, CDPG, FSID, sarthak1@iisc.ac.in
Swarup E., CDPG, FSID, swarupe@iisc.ac.in
Rakshit Ramesh, CDPG, FSID, rrakshit@iisc.ac.in

Image source: Internet

Course Project

- Compulsory!
- Marks: 20%
- Team size: 3-4
- Duration: ~6 weeks (Oct 15th to Nov 30th)
- Initial proposal: Oct 11th
  - Team formation (choose among yourself)
  - Select project topic/domain and/or datasets
- Final project proposal: Oct 18th
  - Detailed information: Problem definition, dataset(s), proposed methodology, and implementation plan.
  - Submission of slides (use the following slides)
- Checkpoints
  - First: Completion of data preparation and EDA (5%)
  - Second: Completion of model development and validation (5%)
  - Final: Final report, project presentation and demonstration (5%)
  - Peer feedback: 5%

Problem Definition

Background

- The Nigerian banking industry reported **₦17.67 billion in fraud losses in 2023**, a 23% increase from 2022, across over 95,000 cases. Mobile and social-engineering-based frauds dominate, highlighting the growing sophistication of cyber-criminals. Traditional rule-based systems struggle with evolving fraud patterns and high false positive rates.

Importance

- Fraud detection accuracy directly impacts customer trust, regulatory compliance, and financial stability. With Lagos state accounting for 48% of fraud cases, there's urgent need for detection systems that can adapt to Nigerian banking patterns, reduce manual intervention, and provide real-time protection

Objectives

- Build an ML pipeline to detect fraudulent transactions within the Nigerian banking ecosystem using the NIBSS Fraud Synthetic Dataset.
- Achieve **F1-score $\geq$ 0.90** and **AUC-ROC $\geq$ 0.95** while maintaining a **false-positive rate < 0.1%**.
- Deliver model interpretability for compliance and stakeholder transparency.

Role of Data Science

- Machine-learning models can learn hidden transaction-behavior patterns and temporal anomalies that rule-based systems overlook. By combining **ensemble learning** and **deep learning**, we can detect new fraud types, reduce manual investigation overhead, and provide interpretable insights for auditors and compliance teams.

Data Collection and Preparation

Data Sources

- **Primary:** 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 analysis of real-world data may require additional anonymization steps that are beyond the scope of this course.

Data Description

- **Size:** 1M records
- **Features:** Transaction Amount, Type, Balance, Channel, Customer Demographics, Merchant Category, Risk Score, Device Info, Location, Transaction Frequency, Account Age, Previous Fraud History, Social Engineering Indicators, and target label *Is Fraud*.
- **Format:** CSV; numeric + categorical + temporal attributes.

Preprocessing Steps

- Missing-value imputation and outlier removal using domain-specific heuristics
- Encoding of categorical variables; normalization of numeric fields
- Temporal feature engineering (hour, day, month, weekday/weekend, seasonality)
- Derived behavioral features: spending velocity, geolocation drift, device mismatch
- Handling of class imbalance with SMOTE, and cost-sensitive learning
- Temporal train/test split to avoid data leakage and simulate real-world fraud detection

Proposed Methodology

- **Analytical Framework**
  - **Exploratory Data Analysis:** Descriptive statistics, correlation heatmaps, geospatial and temporal trend visualizations
  - **Baseline Models:** Logistic Regression (L1/L2) and Random Forest with class weighting
  - **Advanced Models:** XGBoost with custom objectives for imbalanced data
  - **Ensemble Stacking:** Meta-model combining top performers to improve robustness
  - **Prototype:** Streamlit-based dashboard for model inference simulation
- **Tools and Technologies**
  - Python (pandas, numpy, scikit-learn, XGBoost, TensorFlow/Keras, matplotlib, seaborn, plotly),
  - Google Colab Pro / AWS for GPU runtime,
  - GitHub for version control.

Implementation Plan

Phase	Key Activities	Timeline
Week 1	Dataset acquisition, exploratory scan, Nigerian fraud cases literature review	Oct 15–21
Week 2	Data cleaning, preprocessing, feature engineering → *Checkpoint 1*	Oct 22–28
Week 3	Baseline models + imbalance handling; metric framework setup	Oct 29–Nov 4
Week 4	Advanced ensemble and neural models; hyperparameter optimization → *Checkpoint 2*	Nov 5–11
Week 5	Explainability analysis, bias and compliance checks, comparative evaluation	Nov 12–18
Week 6	Streamlit dashboard prototype, integration, final report & presentation	Nov 19–30

Challenges and Risks

Risk	Mitigation
Severe class imbalance	Apply SMOTE variants and cost-sensitive learning methods
Regulatory inexplicability	Integrate SHAP reports and bias detection tests
Computational constraints	Use Colab GPU runtime and sub-samples of dataset
Team coordination issues	Weekly syncs, GitHub branching strategy, shared project board

Expected Outcome

- **Deliverables**
  - High-performing fraud-detection model for the Nigerian banking context
  - Comprehensive EDA report highlighting geographical and temporal fraud patterns
  - Dashboard showing feature importance and transaction-level reasoning
  - Deployment simulation demonstrating real-time prediction capability
- **Success Criteria**
  - **Quantitative:** F1 $\geq$ 0.90; AUC $\geq$ 0.95; False Positive Rate < 0.1%
  - **Qualitative:** Insights that can inform bank fraud strategy and compliance processes; evidence of clear, interpretable model behavior

Role and Responsibilities

Student	Responsibilities
Student 1	Acquire dataset, build data pipeline, handle preprocessing, feature engineering, and maintain data dictionary & project documentation.
Student 2	Conduct comprehensive EDA (fraud trends, channels, regions), develop visual analytics and summary dashboards, contribute to feature selection.
Student 3	Implement baseline models (LogReg, RF), apply class imbalance strategies, perform cross-validation, compare metrics, and refine models.
Student 4	Build XGBoost/LightGBM models, conduct interpretability analysis, develop Streamlit prototype, and coordinate final presentation.

Note: All members share responsibility for model evaluation, report writing, and peer review.

Data Science Canvas

What is a Data Science Use Case Canvas?

A template that helps define the project's context, objectives, data, constraints, and deliverables. It's best to fill out the canvas during a brainstorming session with the project's stakeholders.

References:

- DS Canvas: https://github.com/tomalytics/datasciencecanvas
- ML Canvas: https://github.com/louisdorard/machine-learning-canvas
- ML Canvas – Churn prediction: https://github.com/louisdorard/machine-learning-canvas/blob/master/churn.pdf

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**	
Business Case & Value Added Which business case should be analyzed and what added value does it generate?	**Model Selection** Which analysis methods can be considered on the basis of the specific data landscape and the business case?	**Model Requirements** Which model requirements must be complied with in order to obtain a valid model?	**Skills** What skills are needed to provide the data and model development?	**Model Evaluation** Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?	**Data Storytelling** What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?	**Data Selection & Cleansing** Which of the available data is relevant? Do the data have to be cleaned up?	**Data Collection** How and with which methods should additionally required data be collected? What properties has this data to fulfil?
Data Landscape Which data is required for this and which is already available? Which additional data has to be collected?		**Software & Libraries** Which software should be used? Is there already a standard solution? Which libraries are used?				**Data Integration** In which system should the data from different sources be migrated?	**Explorative Data Analysis** Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

Adopted from: https://github.com/tomalytics/datasciencecanvas