# Black Money Transaction Detection

**Group 8**: Kim Khue Nguyen, Aayushi Patel, Marcanthony Solorzano, Matthew Adent
https://github.com/Aayushi015/Project-4-Group-08

---

## Project Overview

In this project, we developed a machine learning based fraud detection system to identify suspicious financial transactions, often referred to as "black money." Our goal was to classify transactions as either legal or illegal using a real world dataset from Kaggle. This process was supported by extensive data exploration, visualization, and deployment through a Flask web application.

---

## 1. Dataset Selection

We used the **Global Black Money Transactions Dataset** from Kaggle:
Dataset Link: Kaggle Dataset

This dataset includes thousands of anonymized global financial transactions labeled with their fraud status. Features include:

- `amount_usd`, `transaction_type`, `industry`, `risk_score`
- Binary and categorical indicators like `shell_co_involved` and `reported_by_authority, source of money`

**Rationale**
Financial fraud is a growing global issue. This dataset provides a valuable opportunity to apply fraud detection and anomaly detection techniques to real-world-like data.

**Limitations**
The dataset is highly imbalanced, with the majority of transactions labeled as legal, which posed many challenges for model training and evaluation.

---

## 2. Data Cleaning, Preprocessing, and Feature Engineering

**Lead:** Kim Khue Nguyen

In our Jupyter notebooks (`Global_Black_Money_Transactions_1.ipynb`, `2.ipynb`), we:

- Cleaned and normalized data (e.g., removed duplicates, handled missing values).
- Label encoded for binary features / Boolean class (ORDINAL Encoding)
- Encoded categorical features (One Hot Encoding)
- Scaled numerical data
- Defined preprocessing pipelines
- Feature selection and correlation analysis
- Converted preprocessed data to a DataFrame
- Split the dataset into training and testing subsets for machine learning

We built a `sklearn` pipeline and exported the finalized model as `model_pipeline.pkl`.

**Visual Idea:** A flowchart from raw data ➜ cleaned data ➜ features ➜ model input.

---

## 3. Tableau Visualizations

**Leads:** Aayushi Patel, Kim Khue Nguyen

We used Tableau to explore and visualize trends in the dataset. Visual elements included:

- Bar charts of legal vs. illegal transactions
- Correlation heatmaps

- Time-series line charts showing fraud trends over time
- Donut charts visualizing predicted transaction probability

**Color Palette Used:** Coolors palette

**Files:** `Global_black_money_transaction.twbx`, `viz_proj4.twbx`

---

## 4. Machine Learning (Model Training & Evaluation)

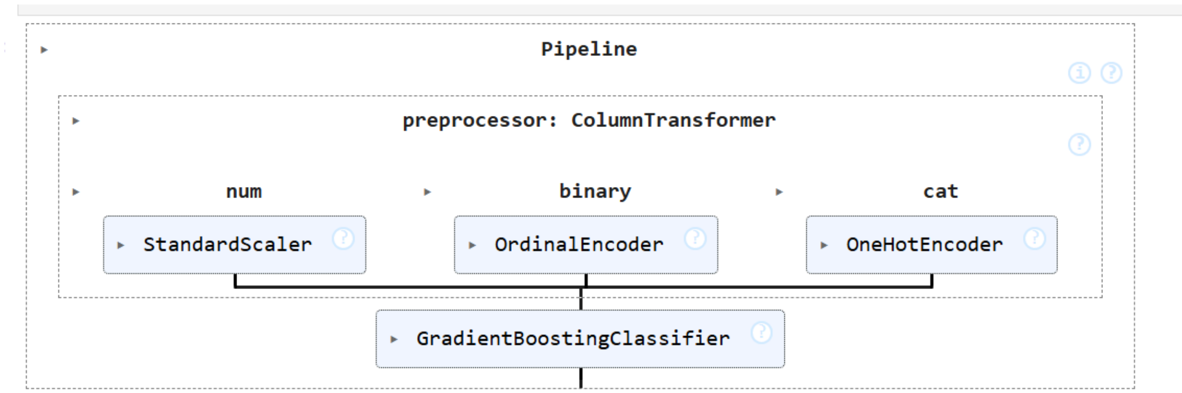
**Leads:** Kim Khue Nguyen, Marcanthony Solorzano

ML approaches to uncovering black money transactions help uncover hidden patterns in large, complex datasets.We evaluated multiple classification algorithms by comparing model performance (accuracy, precision, recall, F1-score), as follows:

| **Model** | **Accuracy** | **Train Metrics AUC** | **Test Metrics AUC** | **Confusion Matrix**<br>**(TP, FP, TN, FN)**<br>**Notes** |
|---|---|---|---|---|
| **Logistic Regression (baseline)** | **70%** | **0.519** | **0.492** | Predicted all illegal transactions, failed to recognize legal transaction |
| **SCV** | **70%** | **0.718** | **0.509** | Overfit<br>Predicted all illegal transactions, failed to recognize legal transaction |
| **KNeighbors** | **73%** | **0.739** | **0.511** | Overfit<br>Incorrectly identified 391 transactions as legal while |

| | | | | |
|---|---|---|---|---|
| | | | | they are illegal => ignored fraud |
| **DecisionTree** | **57%** | **1** | **0.490** | Severe overfit, low accuracy.<br><br>Incorrectly identified 270 transactions as legal on test set while they are illegal => ignored fraud |
| **ExtraTrees** | **62%** | **1** | **0.499** | Severe overfit.<br><br>Incorrectly identified 344 transactions as legal on test set while they are illegal => ignored fraud |
| **Random Forest** | **65%** | **1** | **0.50** | Severe overfit.<br><br>Incorrectly identified 557 transactions as legal on test set while they are illegal => ignored fraud |
| **AdaBoost** | **70%** | **0.521** | **0.480** | Same as logistic regression |
| **Gradient Boosting** | **70%** | **0.682** | **0.494** | Best recall, only model that didn't ignore fraud.<br><br>· TN: 5260: correctly identifies 5,260 illegal transaction<br><br>· FP: 3: Only incorrectly identified 3 illegal transactions as legal. |

The final model selected was **Gradient Boosting**, saved as `black_money_model.h5`.

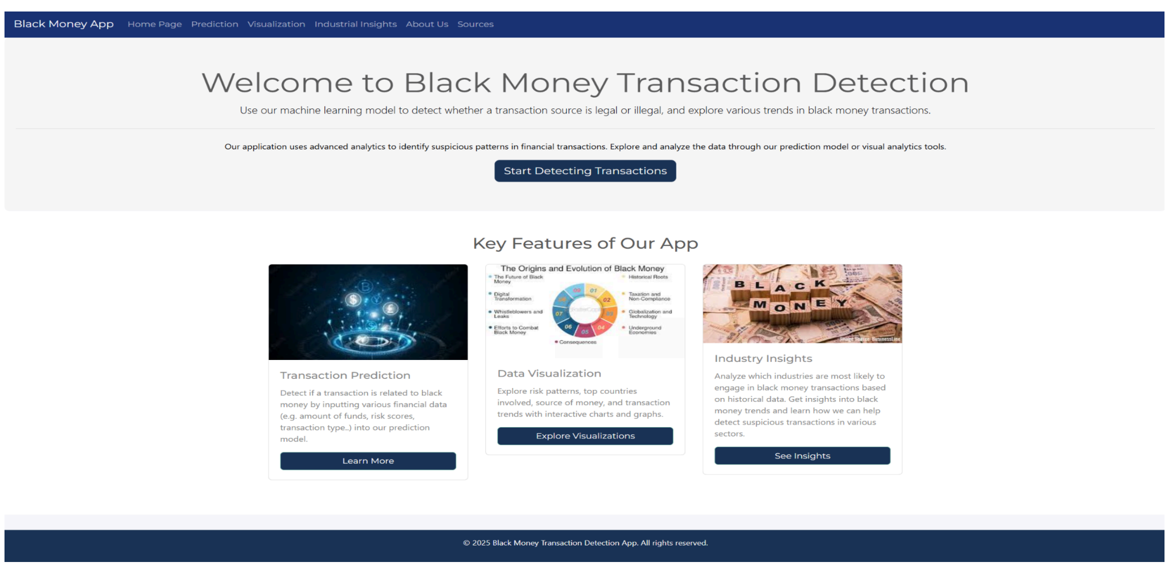


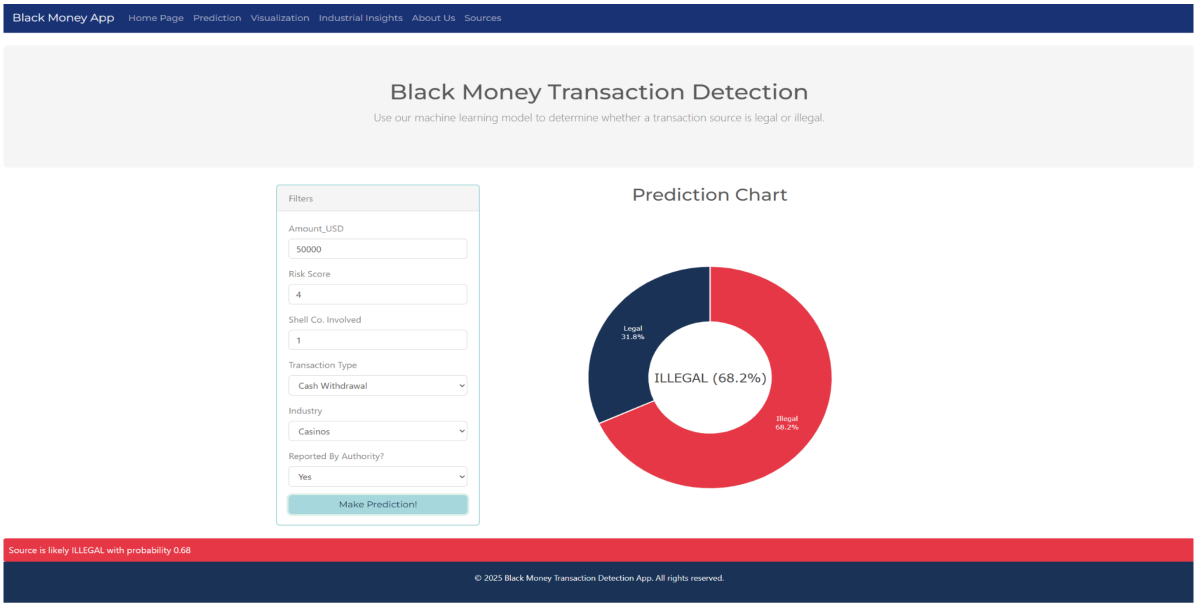

---

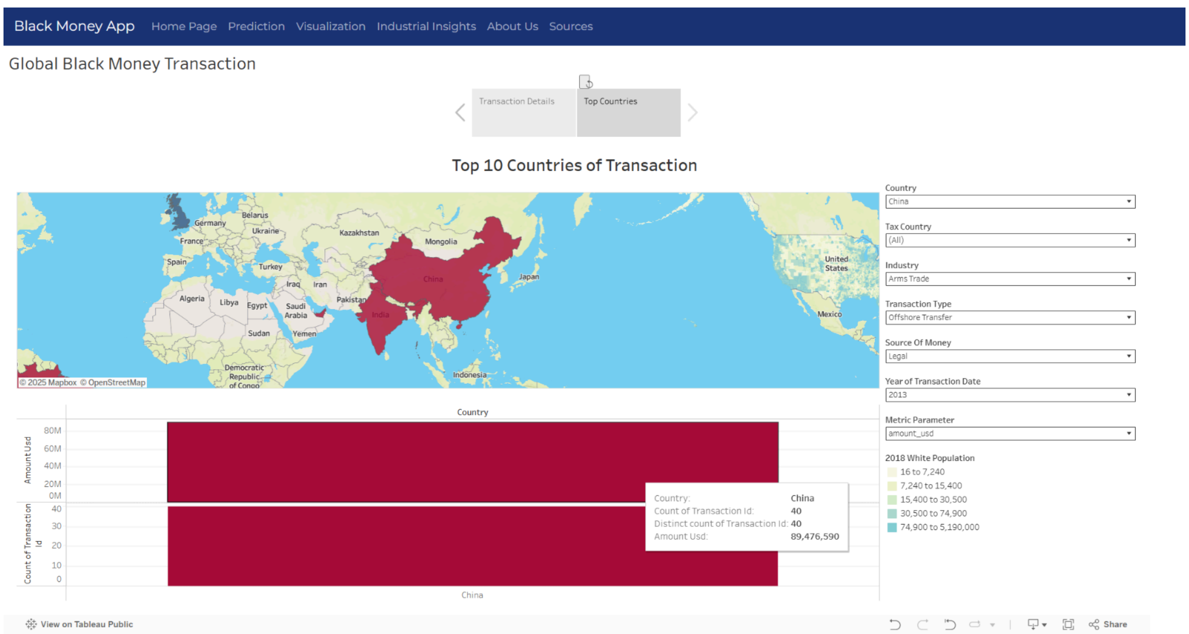
# 5. Flask App Development

**Lead:** Matthew Adent, Kim Khue Nguyen

We built a Flask web application to allow users to input transaction data and receive real-time fraud predictions.






## Key Features:

### Machine Learning Prediction

- Accepts user input for various financial transaction parameters (amount, transaction type, industry, etc.)
- Sends the data to a trained ML model (loaded via `modelHelper.py`)
- Predicts the probability that the transaction source is legal

- Displays results on a dynamic frontend with a donut chart

**Web Interface & Routing**

- Flask app serves multiple HTML pages with navigation:
  - `/` – Landing page with feature highlights
  - `/prediction` – Main prediction interface
  - `/about_us` – Team introduction
  - `/tableau` – Visualizations of transaction trends
  - `/tableau2` – Industry-specific insights
  - `/report` – Writeup
  - `/sources` – Citations and dataset sources

**Data Visualization**

- Tableau visualizations are embedded to explore:
  - Global black money transaction trends
  - Insights by industry and country
  - Correlation heatmaps

**Model Integration**

- Uses a pre-trained ML model (`black_money_model.h5`) for inference
- `modelHelper.py` handles data preprocessing and prediction logic

**Frontend UX**

- Uses Bootstrap (Minty theme) for responsive UI
- JavaScript (`logic.js`) handles client-side prediction requests and visualization rendering
- Custom styles and interactivity via HTML/JS

#### Modular Design

- HTML components separated for reusability (e.g., `navbar.html`)
- Routes and prediction logic are cleanly separated in `app.py` and `modelHelper.py`

---

## 6. Conclusions and Takeaways

- Gradient Boosting performed best of our models, identifying some illegal transactions with low recall
- Tableau made it easy to visualize fraud patterns and feature relationships
- The project taught us about the processes that go into building an end-to-end machine learning workflow from data to deployment

#### Next Steps:

- Improve performance using oversampling
- Experiment with other deep learning approaches
- Deploy the app to the cloud (Heroku, GitHub Pages, etc.)
- Enhance user interface and mobile page (it was not made with mobile in mind)

---

## 7. Data Limitations & Bias

**Small Dataset Size**

- Our dataset is relatively small, which limited our machine learning model.
- With more data, especially recent data, we could train a more accurate model.

**Outdated Data (2013–2014)**

- The dataset is from 2013–2014, which may not reflect current black money transaction patterns.

- Financial crime techniques have evolved significantly in the last decade.

**Real-Life Data, But Limited Scope**

- While the dataset is based on real world transactions, it may not include all the features or scenarios seen in modern financial systems.

---

## 8. Call To Action & Future Work

The analysis of global black money transactions reveals critical vulnerabilities in financial systems, particularly related to shell companies, tax havens, and underreported high risk transactions. Based on these insights, the following actions are recommended:

- Strengthen Regulatory Oversight
  Governments and financial authorities must implement stricter compliance checks
- Empower Analysts with Data Driven Tools
  Equip financial crime units with visual analytics platforms such as Tableau to identify risk patterns in real time and take proactive action.

---

## 9. Future Work

**Integrate Real World Financial Data**
Enhance the dataset with real datasets or open source financial crime data to validate patterns in the synthetic dataset.

**Develop an Early-Warning System**
Create a scoring system or dashboard that flags high risk patterns in real time for investigative use.

**Network Analysis**
Use graph theory to map relationships between people, shell companies, and financial institutions involved in black money flows.

**Create Policy Simulation Tools**
Build models to simulate the impact of certain regulations (ex. limiting shell company usage) on reducing money laundering risk.