# Online Payment Fraud Detection – Analysis Report

---

## Introduction

Online payments have revolutionized the way we transact, but they've also brought along rising cases of fraud. This project is all about building a machine learning system to **detect fraudulent online transactions**, helping financial institutions prevent losses and build trust. The dataset used comes from Kaggle and includes several million transaction records.

---

## Data at Hand

The dataset contains over **6 million online transaction records**, with features describing both the sender and the receiver of the transaction, along with balances before and after the transfer.

**Key Features:**

- **step** – Time unit (1 step = 1 hour)
- **type** – Transaction type (TRANSFER, CASH_OUT, etc.)
- **amount** – Amount of transaction
- **nameOrig** – Sender's name (anonymized)
- **oldbalanceOrg** – Sender's account balance before
- **newbalanceOrig** – Sender's account balance after
- **nameDest** – Receiver's name (anonymized)
- **oldbalanceDest** – Receiver's balance before
- **newbalanceDest** – Receiver's balance after
- **isFraud** – Target variable (1 = Fraud, 0 = Not Fraud)

---

## Exploratory Data Analysis

- The majority of transactions are **legitimate**, with fraud accounting for a small portion.
- **CASH_OUT** and **TRANSFER** are the only transaction types where fraud occurs.
- Fraudulent transactions often involve **zero or inconsistent balances** after the transfer.
- **No missing values** were found in the dataset.

---

## Data Preprocessing

- Removed the **isFlaggedFraud** column (not useful).
- Dropped **nameOrig** and **nameDest** (non-informative for modeling).

- Encoded the **type** column to numerical format.
- Balanced the dataset using resampling techniques due to rarity of fraud cases.

---

**Model Building**

Several models were trained and evaluated:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **XGBoost**
4. **Neural Networks (Keras/TensorFlow)**

Training and testing split: **80/20**

---

**Model Performance Summary**

| Model | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
|---|---|---|---|---|---|
| **Logistic Regression** | 95.5% | 72% | 68% | 70% | 0.85 |
| **Random Forest** | 98.6% | 86% | 84% | 85% | 0.94 |
| **XGBoost** | **99.9%** | **90%** | **88%** | **89%** | **0.97** |
| **Neural Network** | 99.5% | 87% | 85% | 86% | 0.95 |

XGBoost emerged as the best model, offering high accuracy and balance between false positives and false negatives.

---

**Key Insights**

- **Fraud is mostly found in CASH_OUT and TRANSFER transactions.**
- **High-value transfers with balance inconsistencies are key fraud indicators.**
- **Machine learning models can effectively learn these patterns and flag fraud with high precision.**

---

**Conclusion:**

- ML models, especially XGBoost and Neural Networks, can **accurately detect fraud**.
- Proper **data preprocessing and balancing** are essential.
- Fraud often has **distinct balance behaviors** that models can pick up on.