**Portfolio Project Report**

**User Segmentation on PaySim Mobile Money Transaction Data**

**👤 Author: *Peace Igwe***

**📅 Date: *12/10/2023***

**1. Project Overview**

**Goal:**  
This project focuses on applying unsupervised learning techniques to segment users in a mobile money platform using the **PaySim** synthetic transaction dataset. The segmentation helps identify behavioral clusters, detect fraud, and drive customer-focused business strategies.

**Problem Statement:**  
In mobile financial ecosystems, users exhibit diverse transactional behaviors. Grouping users based on these behaviors enables better decision-making in fraud prevention, customer retention, and personalized service delivery.

**2. Dataset Summary**

**Source:** PaySim (synthetic mobile money transaction dataset simulating real-world financial activity)

**Size:** 6.3 million transaction records

**Key Features Used:**

* step – Hourly time index
* type – Type of transaction (TRANSFER, CASH\_OUT, etc.)
* amount – Transaction amount
* oldbalanceOrg, newbalanceOrig – Sender account balance before and after
* oldbalanceDest, newbalanceDest – Recipient account balances
* isFraud, isFlaggedFraud – Fraud indicators

**Data Aggregation:**  
The dataset was aggregated at the **user level**, generating per-user features such as:

* Total and average transaction amounts
* Transaction frequency and type diversity
* Account balance patterns
* Fraud frequency

**3. Methodology**

**Preprocessing**

* Removed system-generated transfers and non-user transactions
* Encoded transaction types using one-hot encoding
* Scaled continuous features with MinMaxScaler
* Created new features: activity duration, average transaction size, fraud ratio per user

**Clustering Approach**

* **Algorithm Used:** K-Means Clustering
* **Optimal Number of Clusters:** 4 (determined using the Elbow Method and Silhouette Score)
* **Dimensionality Reduction for Visualization:** PCA (Principal Component Analysis)

**4. Clustering Results & Interpretation**

After clustering, users were labeled into four distinct segments:

| **Cluster** | **Segment Name** | **% of Users** | **Key Characteristics** |
| --- | --- | --- | --- |
| **0** | Dormant | ~42% | Minimal activity, low transaction frequency |
| **1** | Frequent Users | ~25% | Regular users with medium transaction volumes |
| **2** | High Spenders | ~18% | Infrequent but high-value transactions |
| **3** | VIP Users | ~15% | High-frequency and high-value transactions |

**Cluster Descriptions**

**Cluster 0 – Dormant**

* Users show minimal transaction activity
* Low to no fraud exposure
* Opportunity for reactivation or pruning strategies

**Cluster 1 – Frequent Users**

* Active users with steady transaction patterns
* Mostly small to medium payments and transfers
* Suitable for loyalty campaigns and financial service upsells

**Cluster 2 – High Spenders**

* Few transactions but with very large amounts
* Medium fraud risk; requires monitoring for money laundering
* Ideal for high-value personalized offerings

**Cluster 3 – VIP Users**

* Business-like behavior: frequent + high-value activity
* High fraud risk; prime candidates for fraud detection efforts
* Justifies premium service tiers and account monitoring tools

**5. Business Impact**

**Fraud Risk Management**

* Clusters 2 and 3 identified as higher-risk groups
* Targeted surveillance policies and anomaly detection models planned

**Customer Segmentation for Marketing**

* Dormant users (Cluster 0): Flagged for re-engagement campaigns
* Frequent Users (Cluster 1): Eligible for referral and retention programs
* VIP Users: Qualify for premium product offerings

**Strategic Recommendations**

* Use segmentation in real-time fraud scoring models
* Personalize user journeys and notifications based on cluster profile
* Extend the clustering pipeline with time-series behavioral changes

**6. Tools & Technologies**

* **Languages:** Python
* **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
* **Methods:** K-Means Clustering, PCA, Feature Engineering, Data Aggregation
* **Environment:** Google Colab

**7. Conclusion**

This project demonstrates how unsupervised learning can uncover hidden user patterns in financial transaction data. By mapping user segments into Dormant, Frequent, High-Spending, and VIP groups, we provide a foundation for:

* Smarter fraud detection
* Targeted user engagement
* Data-driven financial service personalization

**8. Future Work**

* Integrate time-based features (session behavior, frequency over weeks)
* Combine clustering with supervised fraud detection models
* Deploy clustering in a real-time streaming context (Kafka + Spark)