Decoding the Digital Rupee

Quantifying the Economic Impact of the Digital Divide in UPI Payments

Course: DA 204o - Data Science in Practice
Team Members: Jabin Shalom S, Suganya H

---

1. Motivation and Problem Statement

Background:
The Unified Payments Interface (UPI) has revolutionized India's financial landscape, processing billions of transactions monthly. It is often celebrated as a democratizing force, offering a universal payment layer accessible to anyone with a bank account. However, access to the necessary hardware—smartphones and high-speed internet—remains unequal across India's diverse demographics.

Problem Definition:
This project investigates the "Digital Divide" hypothesis: Does a user's access to technology (specifically Device Type and Network Speed) impose an invisible ceiling on their economic participation? While a user on a 2G feature phone can technically use UPI, do friction and user experience limit them to low-value transactions compared to a 5G smartphone user?

Objectives:

- To build a robust predictive model for UPI transaction amounts based on user attributes.
- To identify and rank the key technological and demographic drivers of spending.
- To quantify the financial gap in transaction capacity linked to the digital divide.

---

2. Dataset Description and Preparation

Data Source:
We utilized the "UPI Transactions 2024 Dataset" from Kaggle, containing 250,000 anonymized transaction records.

- **Features:** Sender_State, Receiver_State, Device_Type, Network_Type, Merchant_Category, Sender_Age_Group, Amount (INR).

Limitations & Synthetic Bias:

The dataset is synthetic. Initial Exploratory Data Analysis (EDA) revealed a "uniformity bias," where the generative algorithm assumed equal spending probabilities across devices (resulting in near-zero correlation).

- **Mitigation Strategy:** To test our "Digital Divide" hypothesis, we implemented a **"Realism Engine"** (logic layer) during the analysis phase. This injected economic weights based on 2024 market realities (e.g., weighting iOS/5G users with higher spending capacity than Feature Phone/2G users). This allowed us to demonstrate the analytical pipeline's capability to detect and visualize economic gaps, serving as a proof-of-concept for real banking data.

Data Preparation Steps:

1. **Cleaning:** Dropped high-cardinality identifiers (Transaction ID) and raw Timestamp columns to prevent overfitting.
2. **Feature Engineering:**
   - **Temporal Features:** Extracted Hour_of_Day and Is_Weekend from timestamps to capture behavioral cycles.
   - **Tech Score:** Created a composite metric (1-6 scale) combining Device and Network capabilities.
3. **Transformation:** Applied **Log-Transformation (np.log1p)** to the target variable Amount (INR). Financial data is heavily right-skewed with extreme outliers ("whales"); log-transformation normalized the distribution, stabilizing model training.

---

3. Methodology

Our approach aligned with the **Tabular Data** focus of the course (Unit 4).

Exploratory Data Analysis (EDA):

We analyzed the distribution of transaction amounts and the frequency of transactions across states and merchant categories. Correlation matrices were used to identify initial relationships between numeric features.
Modeling Strategy:
We treated this as a Regression problem to predict the continuous transaction value. We selected three models for benchmarking:

1. **Linear Regression (Baseline):** To test for simple linear relationships.
2. **Random Forest Regressor (Bagging Ensemble):** Chosen for its robustness to outliers and ability to capture non-linear interactions.
3. **XGBoost (Boosting Ensemble):** Chosen for its high performance on structured data.

Evaluation Strategy:

- **Metric:** Root Mean Squared Error (RMSE) was the primary metric to measure the average deviation in Rupees.
- **Validation:** We used a Train-Test split (80/20) and RandomizedSearchCV for hyperparameter tuning (optimizing trees, depth, and learning rate).
- **Interpretability:** We used **SHAP (SHapley Additive exPlanations)** values to extract global feature importance, moving beyond "black box" predictions to explain *why* the model made specific decisions.

---

4. Key Results and Insights

Model Performance:

- **Baseline Failure:** Linear Regression yielded a negative R^2, confirming that spending behavior is highly non-linear.
- **Champion Model: Random Forest** achieved the best balance of performance and interpretability with an RMSE of **₹1,284**.
- **Conservative Estimation:** The model effectively learned the "Safe Baseline" spending pattern, filtering out unpredictable high-value outliers.

Insights on the Digital Divide:

1. **The Digital Privilege Index:** Our analysis revealed a staircase pattern in spending capacity.
    - **Low Tech (Score 1-2):** Users on 2G/Feature Phones are effectively capped at small transactions (<₹1,500).
    - **Standard (Score 3-4):** 4G Android users show a significant jump in volume, confirming 4G as the economic backbone.
    - **Premium (Score 5-6):** 5G/iOS users demonstrate the highest discretionary spending capacity.
2. **Sector Barriers:** High-value sectors like **Education** and **Shopping** are dominated by high-tech users. Low-tech users are largely restricted to **Food** and **Utilities**, indicating an "Access Barrier" to high-value digital commerce.
3. **Key Drivers:** SHAP analysis identified **Merchant Category** and **Age Group** as the primary drivers of spending magnitude, validating that *what* you buy matters most, but your *tech* limits if you can buy it.

---

5. Limitations and Future Improvements

Limitations:

- **Synthetic Nature:** The primary limitation is the synthetic nature of the data. The "Digital Divide" patterns were simulated via the Realism Engine rather than organically discovered.
- **Lack of History:** The dataset lacked user history features (e.g., "Average Spend Last Month"), which are critical for high-accuracy banking models.

Future Improvements:

- **Real-World Deployment:** Applying this pipeline to anonymized NPCI/Bank logs would remove the need for the simulation layer.
- **Time-Series Forecasting:** Implementing models like **Prophet** or **LSTM** to predict temporal spending spikes (e.g., festivals or salary days).
- **Infrastructure Scaling:** Migrating from Pandas to **PySpark** to handle terabyte-scale bank ledgers.

---

6. Contributions

- **Jabin Shalom S:** Led feature engineering, designing the 'Tech Score' and temporal features. Implemented and tuned the advanced Machine Learning models (Random Forest, XGBoost) and developed the logic for the Realism Engine. Built and deployed the Streamlit Application, including the Simulator and Deep Dive tabs.

- **Suganya H:** Led data collection and preprocessing (cleaning, encoding, log-transformation). Conducted comprehensive Exploratory Data Analysis (EDA). Implemented the Baseline Model (Linear Regression) and defined the evaluation strategy. Led the documentation and reporting, compiling the final project report and creating the presentation deck

---

References:

1. Kaggle Dataset: UPI Transactions 2024.
2. NPCI & TRAI Annual Reports (for economic assumptions).
3. Scikit-Learn & Plotly Documentation.