Data Science Canvas

Project: Decoding the Digital Rupee: A Predictive Analysis of India's UPI Spending

Team: Jabin Shalom S, Suganya H

Problem Statement

Business Case & Value Added

Objective: Investigate if technology access (Device/Network) creates a "Digital Divide" limiting UPI economic participation.

Value Added: Quantifies the divide in monetary terms (Rupees).

Model Selection

Methods: Random Forest Regressor (Champion), XGBoost (Runner-up), Linear Regression (Baseline)

Justification: UPI spending behavior is highly non-linear. Tree-based ensemble models (Random Forest/XGBoost) effectively capture complex interactions between demographics and technology, whereas simple linear models failed ($R^2 < 0$)

Model Requirements

Type: Supervised Regression (Predicting a continuous transaction value).

Constraints: Must be capable of handling high-cardinality categorical features (e.g., States, Merchant Categories) efficiently.

Explainability: A "White Box" approach is mandatory to fulfill the project objective of identifying key drivers; techniques like SHAP (Shapley Additive exPlanations) are required to provide interpretable feature importance rankings.

Skills

Data Preparation: Proficiency in feature engineering, data cleaning, and handling missing values.

Modeling: Expertise in implementing and tuning machine learning models, particularly Random Forest, Gradient Boosting (XGBoost), and Linear Regression.

Analysis & Interpretation: Skills in interpreting model results, optimizing hyperparameters, and deriving actionable insights.

Data Visualization: Ability to create effective visualizations for EDA and final reporting.

Execution & Evaluation

Model Evaluation

Quality Indicators:

RMSE (Primary): Measures average prediction error in Rupees. Lower is better.

R^2 (Secondary): Validates the model's ability to capture non-linear spending patterns (Digital Divide signal).

Interpretability: SHAP values rank feature importance to confirm logical drivers (e.g., Device Type).

Real-Time Monitoring:

Current: Not applicable

Data Storytelling

Target Group: Financial Policymakers and Bank Executives requiring actionable insights on financial inclusion and market segmentation.

Communication Strategy:

Visual Proof: Use the "Digital Privilege Index" (Staircase Box Plot) to visually demonstrate that better tech equals higher spending capacity.

Actionable Tools: Deploy the Simulator Tool to show how banks can credit-score unbanked users

Data Collection & Preparation

Data Selection & Cleansing

Selection: Selected relevant features: Sender_State, Device_Type, Network_Type, Merchant_Category, and Amount

Cleansing: Dropped high-cardinality identifiers (e.g., Transaction ID) and raw timestamps to prevent overfitting.

Transformation: Applied Log-Transformation (np.log1p) to the target variable (Amount) to normalize the highly right-skewed financial data.

Data Collection

Method: Download the dataset directly from Kaggle as a CSV file .

Properties: The data must contain row-level transaction details (e.g., Amount, Device Type, Network Type) to allow for granular analysis. Since real banking data is restricted, this synthetic dataset serves as a proxy for the required properties.

			Programming: Proficiency in Python and relevant libraries (Pandas, Scikit-learn).	(one-time static analysis). Future: Required to detect data drift (e.g., inflation, new tech) in a live deployment.	based on tech profiles. Impact Analysis: Use the "Sector Domination" chart to highlight specific economic sectors where low-tech users are excluded.	Engineering: Extracted cyclic features (Hour_of_Day, Is_Weekend) and created the composite Tech_Score to quantify digital access.	
Data Landscape Required Data: Transaction-level logs containing Amount, Timestamp, and user metadata Available Data: "UPI Transactions 2024 Dataset" from Kaggle (Synthetic, 250,000 rows). . Data Gap: Real longitudinal user history is missing in the synthetic set but would be required for a production-grade banking model.		**Software & Libraries** Software: Python (Jupyter Notebook/Google Colab) was used as the primary environment for analysis and model development. Deployment: Streamlit was used to create the interactive web application ("Insight Engine") for demonstration. Libraries: Data Handling: Pandas, NumPy. Visualization: Matplotlib, Seaborn, Plotly. Machine Learning: Scikit-learn, XGBoost. Explainability: SHAP (Shapley Additive exPlanations).				**Data Integration** Data Integration System: The data is loaded into a Python environment (using Pandas within Jupyter Notebook/Google Colab) for processing and analysis. Migration: Since the project relies on a single CSV source file, complex data migration or integration from multiple disparate systems is not required.	**Explorative Data Analysis** Findings: Identified that transaction amounts are highly right-skewed with significant outliers ("Whales"). Correlations: Initial analysis showed weak correlations due to the uniformity of the synthetic data. Key Action: Engineered Tech_Score and Tech_Segment features to expose the latent structure connecting digital infrastructure to spending capacity.