FINSHIELD HACKATHON

Introduction

Access to affordable credit in rural India remains a challenge because traditional credit risk models depend on formal financial histories such as PAN-linked CIBIL scores and regular bank statements which a majority of rural citizens lack. Thus, many capable borrowers remain excluded from mainstream financial services.

Abstract

Our project develops an alternative credit risk assessment model for rural and underbanked populations. Leveraging Government of India's digital infrastructure (PFMS/DBT, BBPS, AEPS, UPI, Account Aggregator), our solution gathers and analyzes a person's government scheme receipts, routine utility bill payments, digital and biometric banking activity, and basic account cash flows. This data is then used to train a transparent machine learning model that accurately predicts the applicant's likelihood of timely loan repayment, empowering banks like Bank of India to responsibly lend where no credit history exists.

Problem Statement

Traditional credit scoring methods are not suitable for rural and underbanked Indians, who often transact in cash and hold minimal official documentation. This results in their systematic exclusion from mainstream lending despite repayment capability. The lack of scalable, privacy-compliant, and transparent alternate models exacerbates the problem.

Our Solution

We propose a streamlined credit-scoring system that only uses data already being generated by rural applicants and digital infrastructure already present in Bank of India. No need for PAN cards, new vendors, or complicated setups. Just four types of data that together give a good picture of credit risk.

1. Government Subsidy Payments (PFMS/DBT):

Every time an applicant gets PM-KISAN, MGNREGA wages, LPG subsidy, pension or other schemes, we capture the date and amount via PFMS. If payments come regularly and on time each quarter, it shows income stability. If they're delayed or irregular, it might signal income uncertainty.

2. Utility Bill Payments (BBPS):

Rural users pay electricity, water, mobile recharge, and LPG bills through BBPS. We use their payment history especially how often they pay on or before the due date to measure their financial discipline. The data comes as BBPS CSVs, already flowing into BoI's SFTP.

3. Digital Transactions (AEPS & UPI):

AEPS withdrawals and UPI transfers tell us how active a user is with digital banking. More

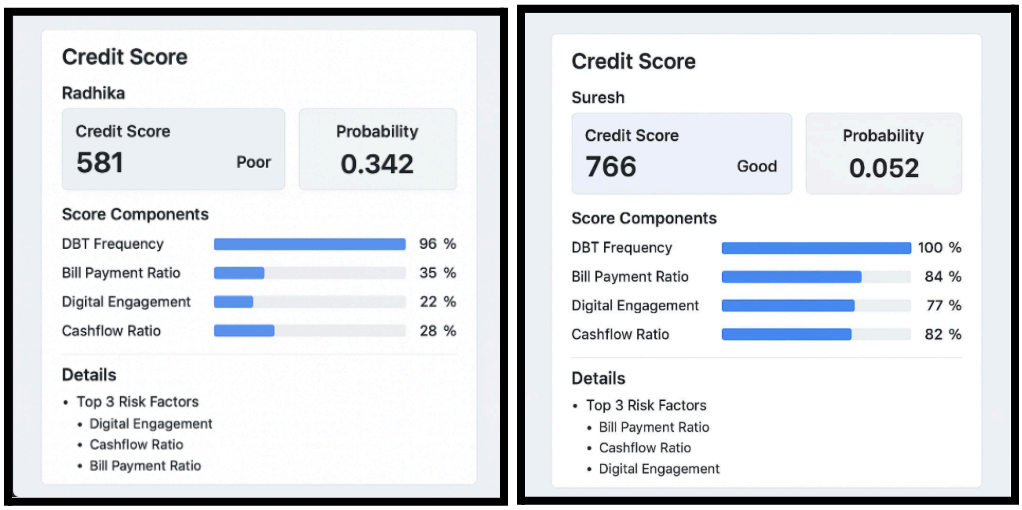
transactions, especially timely ones, reflect trust in formal banking and basic financial literacy. The data is available internally from HDFS logs

4 Cash-Flow Summaries (Account Aggregator):

With the user's consent, we pull monthly deposit and withdrawal totals through the Account Aggregator system. If the deposit-to-withdrawal ratio stays above 1.0, it shows surplus liquidity and low default risk. Consent and data fetch are handled via the AA JSON API

Together, these four inputs subsidies, utility payments, digital transaction count, and cash-flow summaries give us insights into a person's income reliability, repayment discipline, banking engagement, and savings behavior. And since all of this data is already inside BoI's ecosystem, no extra integrations or contracts are needed.

EXPECTED OUTCOME:



Implementation of the Model

We designed a credit scoring model that runs entirely on Bank of India's existing infrastructure using open-source Python libraries no PAN, no third-party vendors, and no new data integrations.

Step 1: Environment Setup

Set up a Python environment (e.g., using Anaconda) and install: pandas (for table data manipulation), numpy (numerical ops), requests (API calls), pyarrow (reading Parquet files), scikit-learn (ML preprocessing), imbalanced-learn (for SMOTE to balance rare defaulter class), XGBoost (main model), and shap (feature explainability). This stack keeps the pipeline fast, simple, and reproducible.

Step 2: Data Ingestion

We collect user-level financial data from four official data sources already flowing through BoI systems and convert them into features.

A. PFMS/DBT (Government Subsidy Payments): Data comes from PFMS API by passing Aadhaar ID to receive a JSON of past subsidies like PM-KISAN, MGNREGA, LPG, and pensions. Using requests and pandas.json_normalize, we extract the payment date, amount, and scheme. Features like dbt_count (frequency of payments) and dbt_avg_delay(irregularity in timing) are calculated.

B. BBPS (Utility Bill Payments): BBPS transactions (electricity, water, LPG, mobile) are available as CSVs downloaded via SFTP. These are loaded using pandas.read_csv, and we compute bill_on_time_ratio, the percentage of bills paid on or before the due date. This feature reflects financial discipline.

C. AEPS & UPI Transactions: Transaction logs from NPCI for biometric (AEPS) and digital (UPI) payments are stored internally in Parquet format (HDFS). Using pyarrow.parquet.read_table, we load only relevant columns like user ID, transaction type, amount, and date. From this, we calculate aeps_txn_count, upi_txn_count, and average amount per user over six months, which show digital engagement.

D. Account Aggregator (AA): With customer consent, AA APIs provide monthly deposit and withdrawal summaries. Using requests and pandas.json_normalize, we compute the

deposit_withdrawal_ratio (total deposits ÷ total withdrawals per month), averaged over the year. It helps assess liquidity. From metadata in the same feed, we extract Aadhaar linkage duration and address history to form the geo_stability variable.

Step 3: Feature Calculation and Model Use

We compute five final features per applicant: dbt_count_qtr (scheduled government credits per quarter), bill_on_time_ratio (utility payment behavior), aeps_upi_txn (AEPS + UPI count), deposit_withdrawal_ratio (cash-flow health), and geo_stability (location stability). All features are scaled between 0–1 using min-max normalization.

The final score is computed using a weighted average:

score = 0.25 * DBT + 0.20 * BillPay + 0.20 * AEPS_UPI + 0.20 * CashFlow + 0.15 * GeoStability.

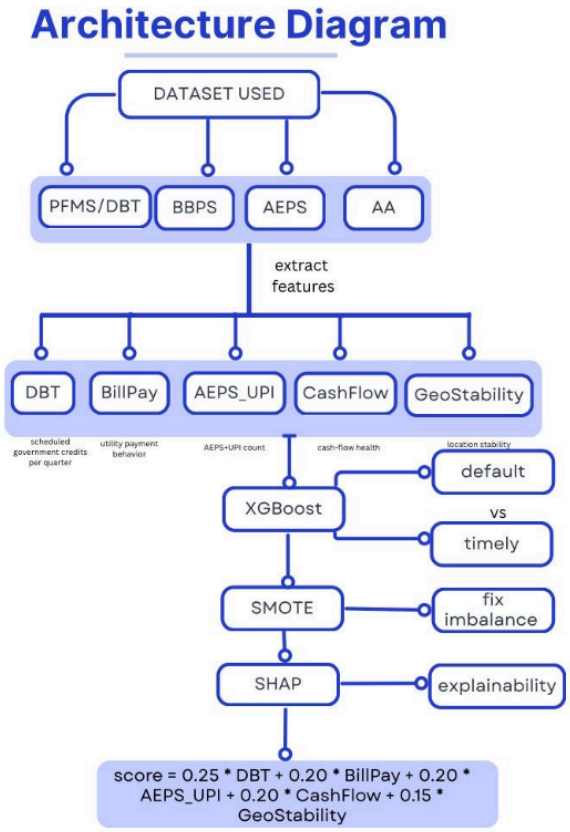
This score is passed to an XGBoost model trained on past labeled data (defaults vs timely repayment). To fix class imbalance (defaults are rare), SMOTE from imbalanced-learn is applied. SHAP is used post-training to explain which features affected individual scores the most, ensuring interpretability.

Why These Libraries?

requests simplifies API calls, pandas and pyarrow handle large structured data efficiently, and scikit-learn + xgboost form a robust ML stack. shap ensures explainability, while imbalanced-learn handles rare default cases effectively, making the pipeline end-to-end reliable.

Why No New Collaborations Needed

All four data sources PFMS (for DBT), BBPS (utility payments), AEPS/UPI (NPCI logs), and Account Aggregator are either government-run or RBI-mandated. BoI already receives or has access to these data streams via its existing infrastructure. No new vendor onboarding or external integration is required, ensuring high trust, better security, and user privacy.



Data sets

A. Sandbox / Dummy Data (for prototyping)

- **PaySim Mobile-Money Dataset** (mimics AEPS patterns):
  Kaggle - PaySim Simulation
- **Synthetic Credit Score Dataset** (10K records, mock profiles with scores):
  Kaggle - Synthetic Credit Scores
- **UPI Transaction Logs** (user-level payment patterns):
  GitHub - UPI Logs

- **PhonePe Pulse District-level UPI Stats** (regional UPI penetration and volume): Kaggle - PhonePe Pulse

B. Production Feeds (Actual Data Sources for Model)

- **PFMS (DBT Inflow Data):** –DBT schemes, amount, frequency, timestamps: https://pfms.nic.in/api/v1/payments
- **BBPS Logs (Utility Payments):** Electricity, water, DTH, LPG payments with timestamps, service type, amount:
  Structured as bbps/YYYY/MM/DD.csv
- **AEPS (Aadhaar Transactions):** Cash withdrawal/deposit logs via micro-ATMs, timestamps, amount, device ID:
  From internal HDFS logs hdfs://bank/aeps_logs/
- **UPI Logs (P2P & Merchant Transactions):** UPI spend/inflow, transaction category, frequency, merchant tags:
  hdfs://bank/upi_logs/
- **Account Aggregator (Bank + Recharge History):** Bank inflow/outflow, EMI, savings, telecom recharge & data pack payments:
  Sahamati /fiu/consent/{customerId}/fetch

1. Technology Stack

- **Python 3.10**
- **Pandas, NumPy** – Data manipulation and numerical computations
- **Scikit-learn** – Preprocessing, model evaluation, and ML utilities
- **XGBoost** – Main model for credit risk classification
- **imbalanced-learn (SMOTE)** – Handling class imbalance by oversampling defaulter class
- **SHAP** – Feature explainability for model transparency
-

2. Expected Outcomes

A. A Credit Scoring Model based on Alternative Data

A trained XGBoost model that predicts the probability of loan payback default using alternative input variables such as:

1. Government subsidy (DBT) behavior
2. Utility bill payment regularity
3. Digital transaction activity (AEPS, UPI)
4. Cash flow balance (deposits vs withdrawals via AA)
5. Geo-stability and Aadhaar linkage duration

The output will be a risk score, ranging from 0 to 100, which would be an indicator of the default probability of a user.

B. A Transparent System for Calculating the Creditworthiness of a Loan Taken

Using SHAP increases the explainability of the result, and it enlightens upon:

1. The reason of a particular score allocated to a user
2. The features behind the credit assigned

Overall, the result is a more trustworthy model.

C. An Inclusive Approach towards Credit Calculation

The model, due to the employment of SMOTE and factual dataset, includes the rural population, with no existing PAN, with an inconsistent banking history and also makes sure to not use just one factor for the CIBIL score prediction, but proves to be multi-faceted in its approach.

D. Uses a Feasibly Deployable Pipeline Inside BoI

Since our datasets make use of source such as PFMS APIs, BBPS CSVs, HDFS logs and AA APIs, it is coherent with Bank of India's current architecture.

E. Impact on the Market

- Enables loans to the under-credited without increasing risk
- Helps meet financial inclusion goals of the RBI & Government of India
- Could improve loan recovery rates by more accurate targeting
- Protects BoI from fraud and ghost applications through data cross-verification

3. Resources

Academic & Policy Papers

Machine Learning for Digital Credit Scoring in Rural Finance -
https://www.mdpi.com/2227-9091/9/11/192
A comprehensive literature review on ML technologies for rural credit scoring, including datasets like satellite imagery, mobile usage, and agronomic surveys.

AI-Driven Innovations in Credit Scoring Models -
https://ijrmp.org/wp-content/uploads/2025/01/in_ijrmp_Oct_2024_GC240198-AP07_AI-Driven-Innovations-in-Credit-Scoring-Models-for-Financial-Institutions-62-79.pdf
Explores how AI enhances fairness, accuracy, and inclusivity in credit scoring, with emphasis on explainable AI and ethical considerations.

Alternative Scoring Models for the Financially Underserved -
https://ifmrlead.org/wp-content/uploads/2025/01/Alternative-Scoring-Models-on-Access-To-Credit-Among-the-Financially-Underserved_Ecosystem-Snapshot_November-2024.pdf
A snapshot of fintech innovations and alternate data usage in credit scoring, especially for rural and low-income populations.

Institutional Reports & Guidelines

NABARD Guidance Note on Credit Risk Management -
https://www.nabard.org/CircularPage.aspx?cid=504&id=16418
Offers a detailed framework for credit risk governance in rural banks, including CAMELSC-based supervision and risk appetite triggers.

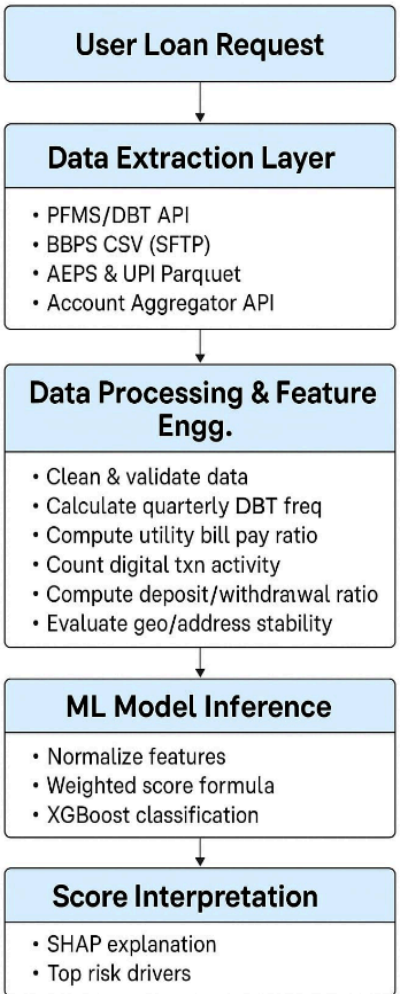
NABARD Rural Financial Inclusion Survey (NAFIS) -
https://static.pib.gov.in/WriteReadData/specificdocs/documents/2024/oct/doc20241010414001.pdf
Provides insights into income, savings, KCC penetration, and insurance coverage among rural households.

Grant Thornton Report on Financial Inclusion in Rural India -
https://www.grantthornton.in/insights/articles/financial-inclusion-in-rural-india/
Covers public and private sector initiatives, fintech innovations, and challenges in rural banking infrastructure.

Datasets for Model Development

Spectral Labs Credit Scoring Dataset -
https://huggingface.co/datasets/spectrallabs/credit-scoring-training-dataset
Tabular dataset with over 400K records including wallet activity, transaction counts, and gas fees, useful for behavioral modeling.

Credit Score Classification Project on GitHub -
https://github.com/saurabhnirwan11/Credit-score-classification
Includes training notebooks, Flask web app, and sample datasets for credit score prediction using ML.



4. Conclusion

India's rural borrowers have long been excluded from formal credit systems due to the absence of standardized financial records. This research presents a transformative alternative - an inclusive, data-driven credit assessment model built entirely on publicly mandated infrastructure already integrated with Bank of India. By leveraging PFMS/DBT subsidies, BBPS utility payments, AEPS/UPI digital transactions, and Account Aggregator cash-flow summaries, our model captures repayment potential through everyday financial activity rather than formal documentation.

Implemented using open-source Python tools and trained via explainable machine learning, our solution ensures transparency, data privacy, and scalability without needing new collaborations or external vendors. Most critically, it empowers financial institutions to extend credit

responsibly to capable borrowers who have been systematically overlooked by legacy scoring models.

This approach doesn't just enhance credit accessibility - it redefines financial inclusion for rural India, aligning perfectly with the vision of a digitally empowered economy.