वित्तीय सेवाएं विभाग
DEPARTMENT OF
FINANCIAL SERVICES

FINNOVATION

Detect Loan Defaulters, Trace their Location & Track them in Digital Ecosystem

TEAM NAME : PATHFINDERS

PRANJAL SONI | KAHAAN SONI | UTKARSH SHARMA
PULKIT GARG | AKASH IYER | LAKSHITA AGARWALLA

LINK TO GITHUB: HTTPS://GITHUB.COM/PATHFINDERS-SBI/SBI-HACK

"In a world of imperfect data, true innovation lies in creating clarity from chaos."

PATH FINDERS

UNDERSTANDING THE PROBLEM STATEMENT

SBI

SMA-2: QUIET PRECURSOR TO NPAS

Many future NPAs originate from SMA-2 accounts that appear normal but are at high risk. Traditional systems fail to flag them in timeBanks struggle with early identification of high-risk borrowers.

MISSED INTERVENTION COSTS BANKS

Without early warning, banks lose the chance to recover before full default — leading to higher NPA provisioning and credit losses

THE CURRENT SCENARIO

DEFAULTERS DISAPPEAR FAST

Over 60% of defaulters go untraceable within 90 days. Lack of contact data cripples recovery, making traceability a key challenge

DEFAULTS ARE HARDER TO DETECT

Subtle behavior patterns and data noise make defaulters blend in. This demands robust, interpretable AI models to spot early risk

PROBLEM STATEMENT

Predict loan accounts likely to become NPAs (SMA-2) through TARGET classification (1 = Defaulter, 0 = Non-Defaulter), and ethically trace the exact location and digital footprint of defaulters

STRATEGIC RELEVANCE

BEHAVIORAL VALIDATION & DIGITAL FOOTPRINT

ANOMALY & RISK SIGNAL DETECTION

NETWORK & RELATIONSHIP ANALYSIS

CREDITWORTHINESS & TRUST SCORING

NPA JOURNEY

1. **Standard Asset** (EMI UNPAID ≤30 DAYS)
2. **SMA-0 (0–30 days)** (OVERDUE CONTINUES PAST 30 DAYS)
3. **SMA-1 (31–60 days)** (OVERDUE CONTINUES PAST 60 DAYS)
4. **SMA-2 (61–90 days)** (OVERDUE >90 DAYS)
5. **NPA (Default)**

PATH FINDERS

APPLICATION OVERVIEW

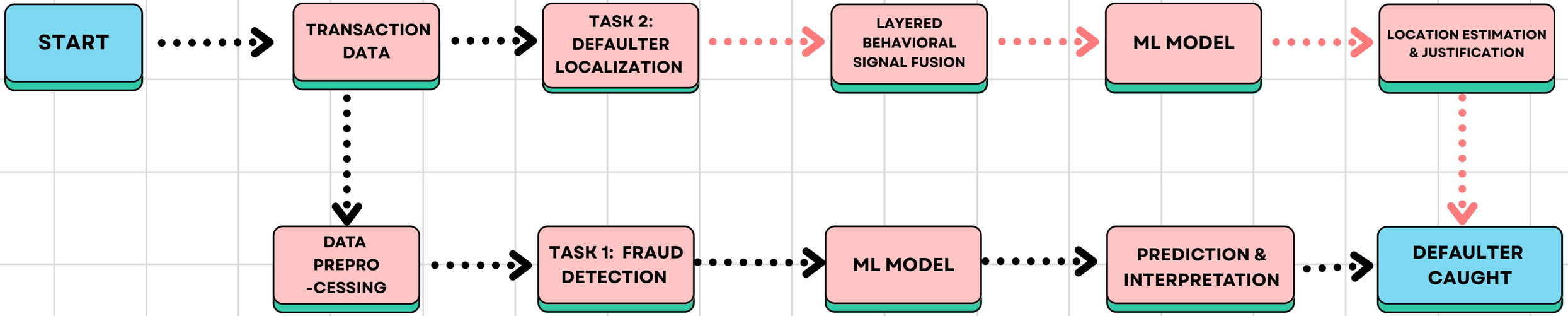
SBI

TASK - 1 : DEFAULT DETECTION

- Identify anomalous transactions and accounts by modeling typical financial behavior patterns.
- Handle class imbalance due to fewer fraud cases compared to normal ones.
- Account for adversarial scenarios where fraudsters may try to evade detection.
- Ensure model interpretability to understand and explain predictions.

TASK - 2 DEFAULTER LOCALIZATION

- Develop ML models to estimate the last known location of loan defaulters.
- Use Layered Behavioral Signal Fusion , combining data from social media activity, ATM, transaction IPs, to estimate defaulters' last known locations.
- Apply spatio-temporal clustering to identify habitual movement patterns of individuals.
- Build a probabilistic model to rank potential last-seen locations based on data reliability and recency.



PATH FINDERS

OVERVIEW OF THE DATASET

SBI

DATA COLUMNS	COLUMN NAME	DESCRIPTION
Customer Demographics & Profile	• AGE: • KYC_SCR: • KYC_FLG, EKYC_FLG, UID_FLG, INB_FLG: • LOCKER_HLDR_IND: • SI_FLG:	Age of the customer KYC Score (Verification/Trustworthiness) KYC-related flags Whether the customer has a locker Standing Instruction flag (autopay indicator)
Account & Loan Metadata	• ACCT_AGE: • LIMIT: • LOAN_TENURE: • ACCT_RESIDUAL_TENURE: • INSTALAMT: • VINTAGE: • NO_LONS: • ALL_LON_LIMIT: • ALL_LON_OUTS: • ALL_LON_MAX_IRAC:	Age of the loan account Sanctioned credit/loan limit Total loan duration in months Remaining loan tenure Monthly EMI/instalment amount Time since first borrowing relationship Number of active loans Combined credit limit across all loans Combined outstanding balance across loans Highest IRAC (NPA classification)
Outstanding Balances & Repayments	• OUTS: • ONEMNTHOUTSTANGBA etc	Current outstanding amount for the loan Columns ending in OUTSTANGBAL across months
Credit and Debit Activity	• Columns ending in SCR • ONEMNTHCR, TWOMNTHSCR etc • Columns ending in SDR • ONEMNTHSDR, TWOMNTHSDR etc	Credit inflow
Account Utilization & Trends (Averages)	• Columns ending in: • AVGMTD: • AVGQTD: • AVGYTD:	Average Monthly Turnover Daily Average Quarterly Turnover Daily Average Yearly Turnover Daily
Flags & Indicators	• KYC_FLG, EKYC_FLG, INB_FLG, • UID_FLG, LOCKER_HLDR_IND, SI_FLG	Binary flags indicating customer behaviors:

DATA COLUMNS ADDED

BANK PERSONAL DATA

- NAME
- EMAIL
- PHONE
- IMAGE
- ADDRESS

- GEOTAG DATA
- DEVICE FINGERPRINT
- TRANSACTION DATA

PATH FINDERS

AI MODEL USED FOR TRAINING

SBI

STEP-1: DATA CLEANING

Type Inspection

Checked column types and unique values.

Object Conversion

Transformed object columns into numeric format.

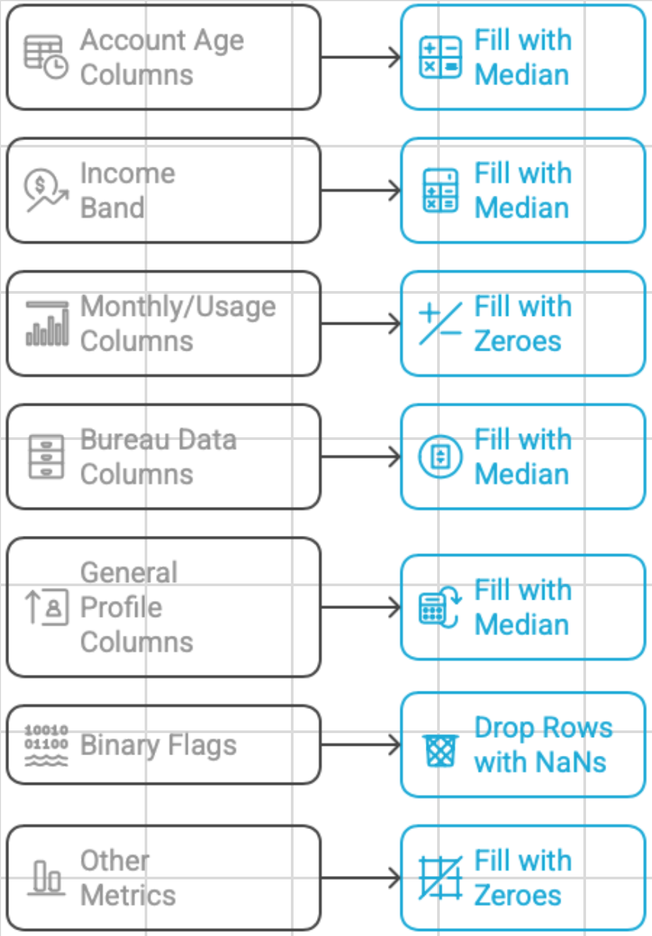
Regex Parsing

Parsed columns using regular expressions.

Ordinal Mapping

Mapped income bands to ordinal integers.

STEP-2: HANDLING MISSING VALUES (NANS)



OVERVIEW 1: DATA TYPE CONVERSION

Converted textual durations like "2yrs 3mon" into numeric months.
Mapped categorical income bands (A–H, EX01–EX05) to ordinal integers.
Transformed all object and boolean columns into numeric format.

OVERVIEW 2: MISSING VALUE HANDLING

Imputed key columns with median values (e.g., account age, bureau data).
Filled transactional columns with zeroes.
Dropped rows with missing critical flags (e.g., UID, KYC).
Ensured complete and clean numeric dataset for modeling.

PATH FINDERS

OVERVIEW OF THE DATASET

SBI

STEP-3: FEATURE ENGINEERING

BEHAVIORAL RATIOS

- UTILIZATION_RATIO = OUTS / (LIMIT + 1)
- VINTAGE_RATIO = VINTAGE / (AGE + 1)

RESIDUAL_RATIO = ACCT_RESIDUAL_TENURE / (LOAN_TENURE + 1)

CREDIT SCORE TRENDS

- CRIFF_MEAN, CRIFF_STD = MEAN AND STD DEV OF CRIFF SCORES (11 TO 66)
- CRIFF_TREND = CRIFF_11 – CRIFF_66 (EARLY SCORE DROP/RISE)

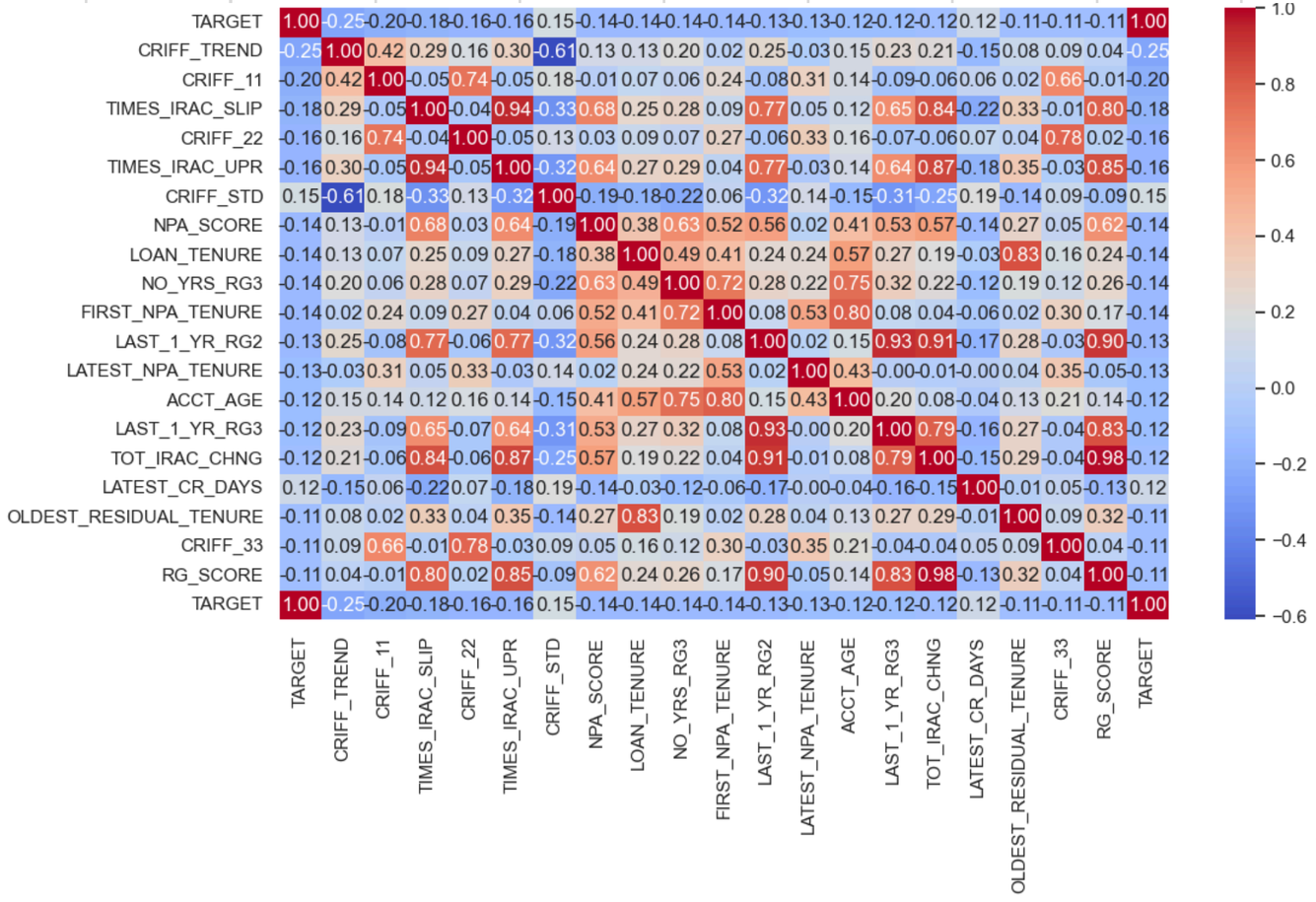
SAVINGS FLOW BEHAVIOR

- AVG_MONTHLY_CREDIT, AVG_MONTHLY_DEBIT = AVG CREDIT/DEBIT ACROSS 12 MONTHS
- CREDIT_DEBIT_RATIO = CREDIT TO DEBIT RATIO ACROSS MONTHS

LOAN BURDEN SIGNALS

- DISBURSE_LIMIT_RATIO = DISBURSED AMOUNT / (SANCTIONED AMOUNT + 1)
- EMI_BURDEN = MONTHLY EMI BURDEN RELATIVE TO DEBIT ACTIVITY

CORRELATION MATRIX

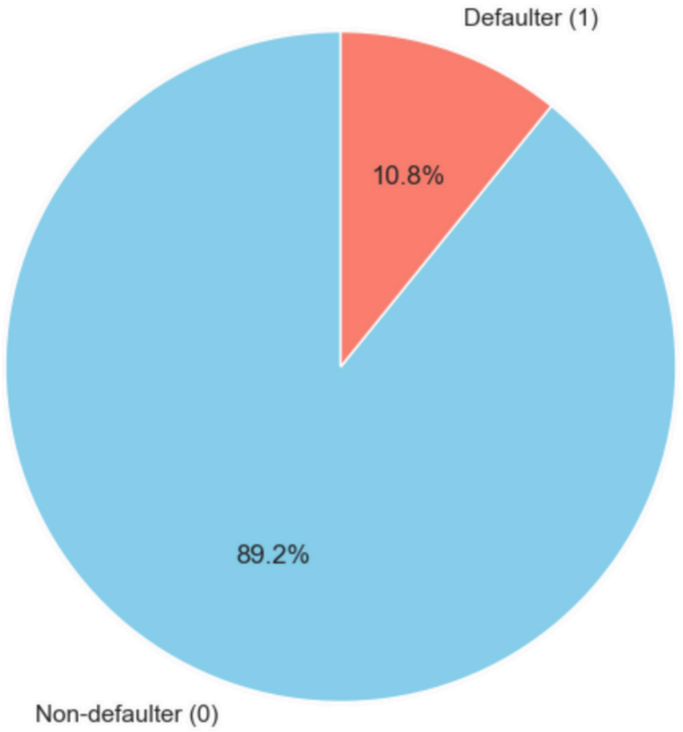


PATH FINDERS

AI MODEL USED FOR TRAINING

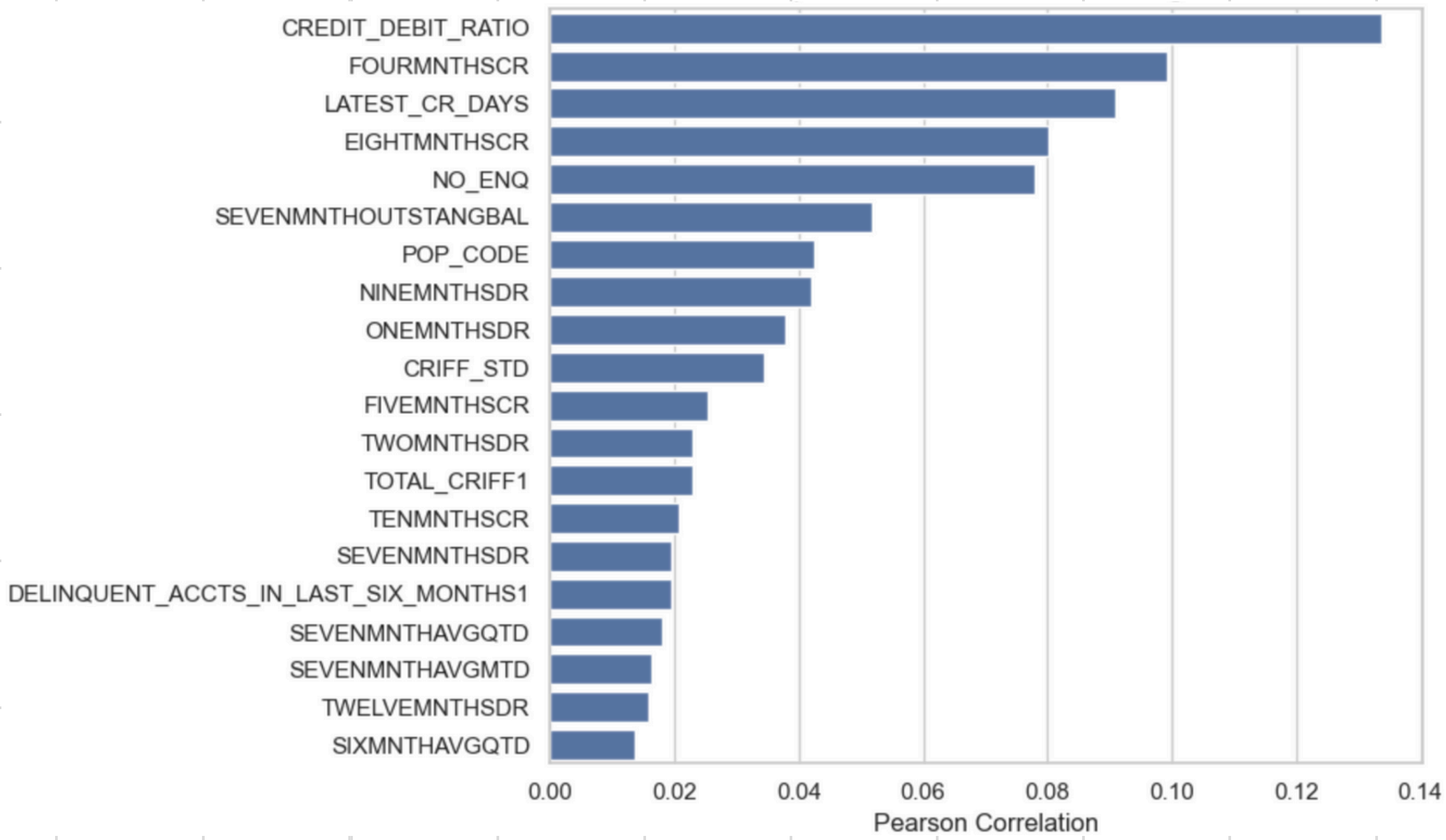
SBI

STEP-4: EXPLORATORY DATA ANALYSIS



SHARE OF TARGET CLASSES IN TRAINING SET

TOP 20 FEATURES CORRELATED WITH TARGET



PATH FINDERS

AI MODEL USED FOR TRAINING

SBI

STEP 5: OUTLIER DETECTION & TREATMENT

- Selected all numeric columns excluding flags, tenure, KYC etc.
- Clipped values to the 1st and 99th percentile range for each numeric feature

STEP 6: CONVERTED BOOLEAN COLUMNS (LIKE FLAGS) TO INTEGERS

- One-hot encoded:

AGREG_GROUP, PRODUCT_TYPE, and TIME_PERIOD

STEP 7: TRAIN-TEST SPLIT

- Used train_test_split with:

test_size = 0.2

stratify = y to preserve class distribution

STEP 8: CLASS IMBALANCE HANDLING

- Used compute_class_weight() to generate
- class_weight_dict for training

STEP 9: NEURAL NETWORK MODEL (KERAS)

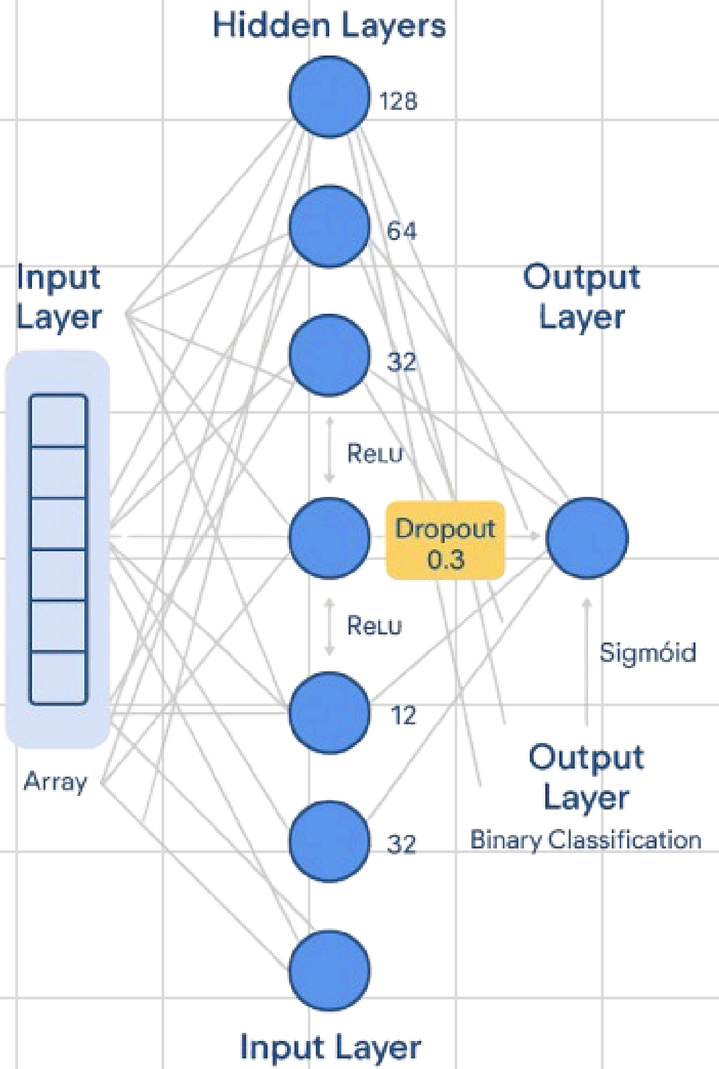
- **Architecture:**

Dense (128 units) + ReLU
→ Dropout (0.3)
Dense (64 units) + ReLU
→ Dropout (0.2)
Dense (32 units) + ReLU

Output:

- Dense (1 unit) + Sigmoid
- Loss: Binary Crossentropy
- Optimizer: Adam

COMPLETE NEURAL NETWORK OVERVIEW



PATH FINDERS

AI MODEL USED FOR TRAINING

SBI

STEP 10: TRAINING LIGHTGBM BASE MODEL

- Trained a **lightgbm (Light Gradient Boosting Machine)** Model as a base model
- LightGBM supports class weight='**balanced**', which automatically adjusts weights to handle imbalance in the data.

STEP 11: TRAINING XGBOOST BASE MODEL

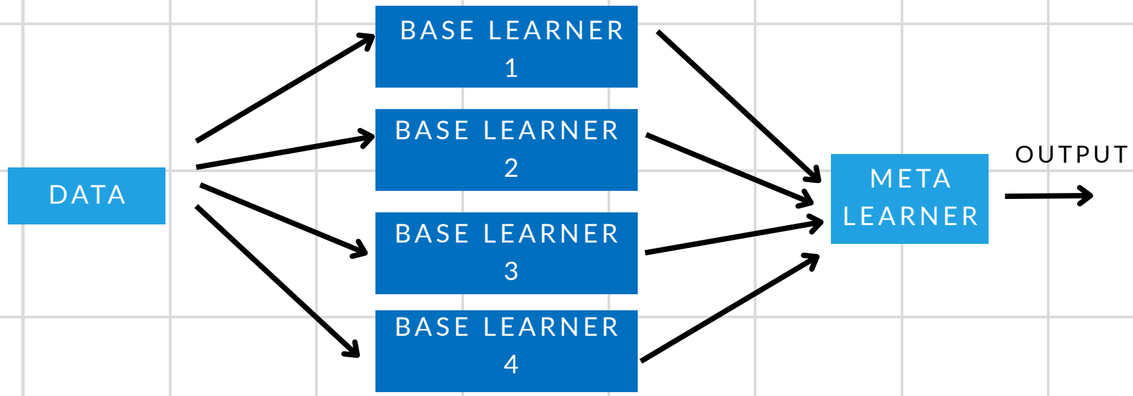
- Trained an **XGBoost** (Extreme Gradient Boosting) model as a base classifier for defaulter prediction.
- Handled class imbalance using **scale_pos_weight,** which boosts the model's sensitivity to the minority (defaulter) class.

STEP 12: ENSEMBLING VIA MANUAL STACKING

- Manual stacking guarantees true **out-of-fold** predictions so the meta-model never sees data the base models were trained on.
- Perform a two-fold stratified split of **X_train_scaled**/**y_train**, training fresh NN and LightGBM models on each fold and predicting probabilities on its hold-out slice

STEP 13: IMPLEMENTING THE META MODEL

- Creating a **Logistic Regression** meta model on those oof predictions
- Logistic Regression offers an interpretable linear blend of the NN and LightGBM scores. Its sigmoid output gives calibrated probabilities, and built-in regularization handles imbalance and prevents overfitting.

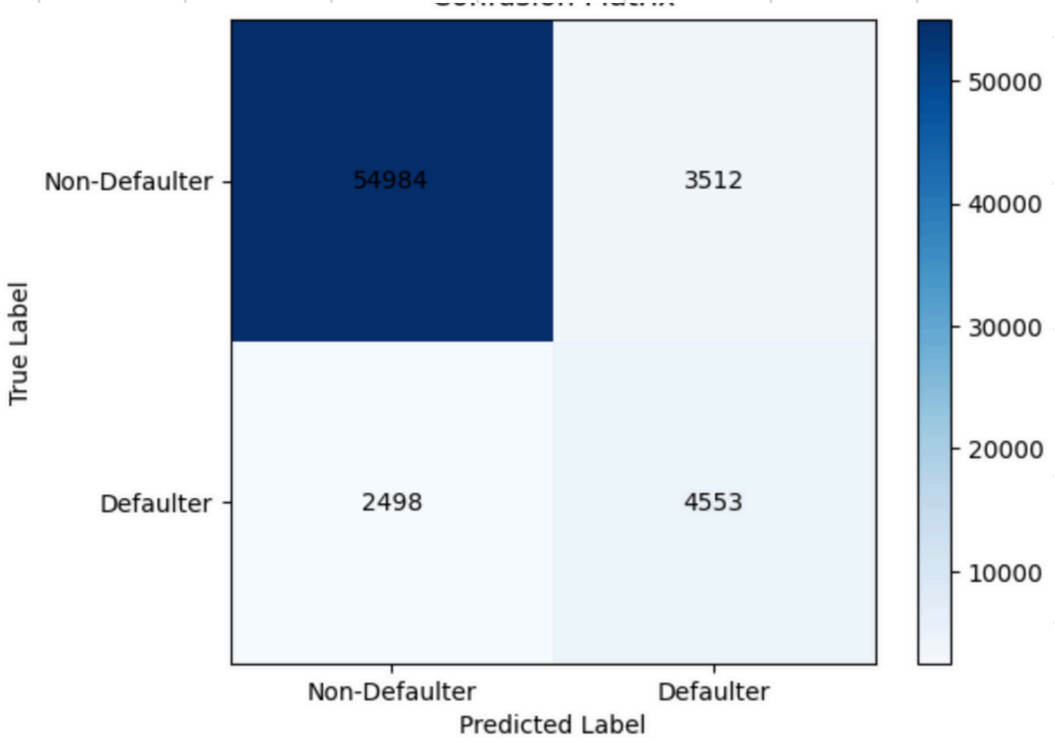


PATH FINDERS

AI MODEL USED FOR TRAINING

SBI

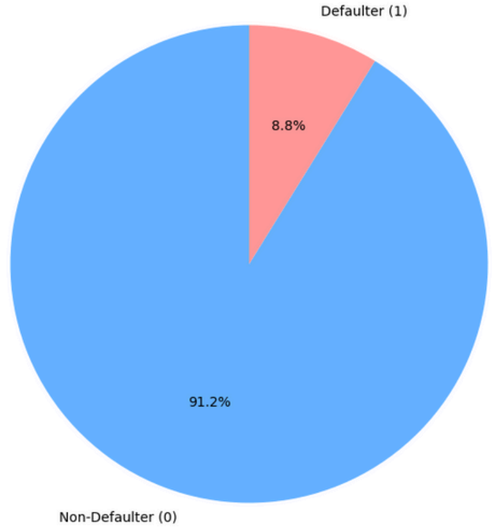
CONFUSION MATRIX



CLASSIFICATION REPORT

	precision	recall	f1-score	support
Non-Defaulter	0.96	0.94	0.95	58496
Defaulter	0.56	0.65	0.60	7051
accuracy			0.91	65547
macro avg	0.76	0.79	0.78	65547
weighted avg	0.91	0.91	0.91	65547

SHARE OF TARGET CLASSES IN PREDICTION DATASET



PATH FINDERS

DETECTING THE LAST TRANSACTIONAL LOCATION

SBI

WE USE A STRUCTURED APPROACH TO FIND THE FINAL LOCATION OF THE DEFAULTERS

DEFAULTER DETECTED

TRANSACTION DATA

OFFLINE ATM TRANSACTION

POP CODE CLUSTERING

LOCATION CLASSIFICATION

IP ADRESS

LOCATION

INTERNET PROVIDER

IP ADRESS

DEFAULTER BANK DATA INFO

DEFAULTER INFO ML MODEL

DIGITAL FOOTPRINTS

SOCIAL MEDIA

LINKEDIN SCRAPPER

TWITTER SCRAPPER

INTERNET DATA FLOATING

MULTILAYER MODEL

DIGITAL PRESENCE

SENTIMENTAL ANALYSIS

RISK SCORE DEFAULTER ANALYSIS

PATH FINDERS

LOCATION & IP TRACKER – FINDMARK

SBI

A critical component of the Defaulter Risk Analysis system, delivering real-time geospatial intelligence to enhance recovery operations.

SYSTEM ARCHITECHTURE

CUSTOM LINK SENT TO DEFAULTER	SBI LANDING PAGE (IFRAME)	PERMISSION FOR GPS LOCATION & IP ADDRESS	FLASK SERVER RECEIVES AND STORES THE DATA	DASHBOARD RENDERS MAPS & METADATA
Specially crafted link that appears to be from SBI	redirects to the SBI landing page for the defaulter to explore.	Browser permission pop up request appears to user	Backend processing and structured data storage	Visualization of captured location and device data

CONSENT & LEGAL COMPLIANCE

All tracking is performed through explicit user consent via browser permission dialog, fulfilling both legal requirements and ethical standards.

PATH FINDERS

LOCATION & IP TRACKER – FINDMARK

SBI

A critical component of the Defaulter Risk Analysis system, delivering real-time geospatial intelligence to enhance recovery operations.

TECHNOLOGY STACK

FRONTEND

HTML + JavaScript (Geolocation API)

User interface and location capture

HOSTING

Serveo / Ngrok / Cloudflare Tunnel

Secure tunneling and access

BACKEND

Python + Flask (via r4ven)

Data processing and storage

LOGGING

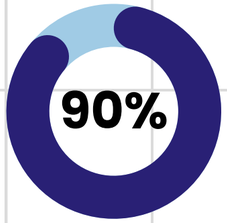
JSON storage (location_log.json)

Structured data persistence

VISUALIZATION

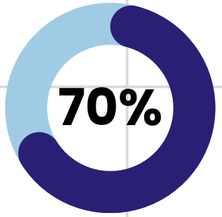
Google Maps embed + Custom Dashboard

Geospatial rendering and analysis



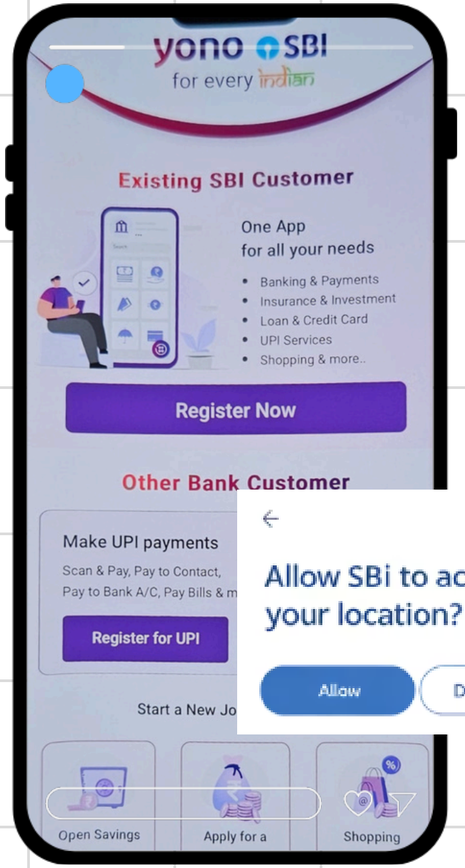
SUCCESS RATE

When users grant permission, location accuracy is within 10 meters for 90% of cases



CONSENT RATE

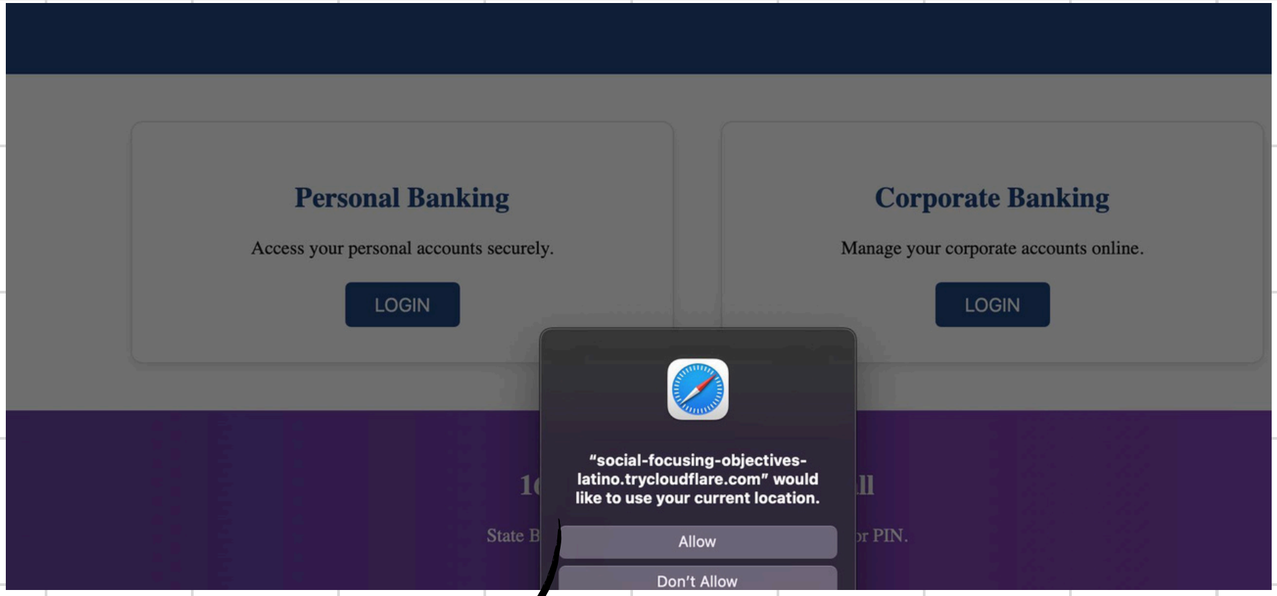
Approximately half of defaulters grant location permission when presented with the SBI interface



PATH FINDERS

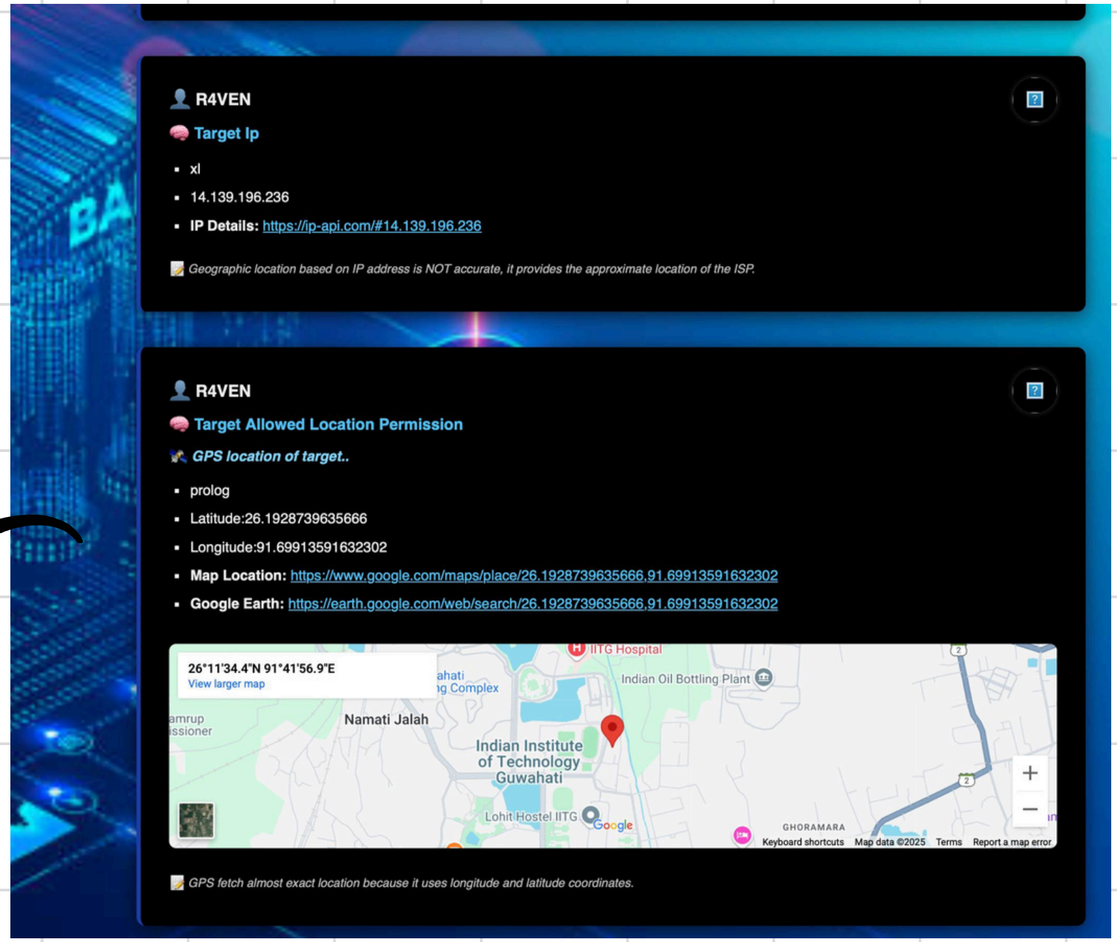
LOCATION & IP TRACKER – FINDMARK

SBI



LINK WILL REDIRECT TO THE SBI LANDING PAGE AND ASKS FOR LOCATION ACCESS FROM THE USER

THE BANK GETS ACCESS TO THE IP DETAILS OF THE DEFAULTER AND AN APPROXIMATE LOCATION ON GOOGLE MAPS



PATH FINDERS

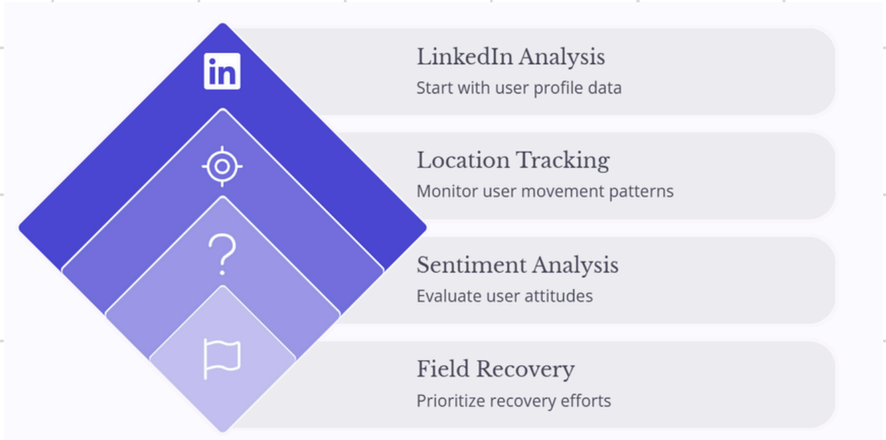
LINKEDIN & TWITTER-BASED FINANCIAL RISK ANALYZER – TRACELINK & TRACETWEET

SBI

A hybrid system to detect financial behavior risks from LinkedIn & Twitter user sentiment using FinBERT, VADER & FastAPI

THE GOAL

- Develop a sentiment-driven risk scoring engine using public LinkedIn & Twitter posts and financial sentiment models.
- Social media, especially LinkedIn, contains rich behavioral signals related to employment, achievements, financial attitude, and risk appetite.



SYSTEM ARCHITECHTURE

FRONTEND

Collects user input: Name, Location, and SerpAPI Key

BACKEND (FASTAPI)

Processes requests, coordinates LinkedIn search, scraping, and sentiment analysis

ANALYSIS ENGINE

Applies FinBERT and VADER models to compute risk scores and categories

PATH FINDERS

TRACE-LINK & TRACE-TWEET

SBI

A hybrid system to detect financial behavior risks from LinkedIn user sentiment using FinBERT, VADER & FastAPI

TECHNOLOGY STACK

WEB FRAMEWORK

FastAPI (Python)

- High-performance, async-capable API framework
- Built-in validation, documentation
- Serves both API endpoints and static frontend

NLP MODELS

FinBERT & VADER

- FinBERT: Financial-domain BERT model
- VADER: Lexicon-based social media sentiment analyzer
- HuggingFace Transformers & NLTK

DATA COLLECTION

SerpAPI & Selenium

- SerpAPI: Google Search API for profile discovery
- Selenium: Browser automation for post scraping
- Chrome WebDriver with persistent user profiles

CALCULATION

$$\text{WEIGHTED\_RISK} = \frac{(\text{POSITIVE}\times 0.2)+(\text{NEUTRAL}\times 0.5)+(\text{NEGATIVE}\times 0.8)}{\text{TOTAL POSTS}}$$

- POSITIVE POSTS = 0.2 (LOWER RISK)
- NEUTRAL POSTS = 0.5 (MODERATE RISK)
- NEGATIVE POSTS = 0.8 (HIGHER RISK)

PATH FINDERS

TRACE-LINK & TRACE-TWEET

SBI

SBI ONLINE

TraceLink

ENTER THE DEFAULTER'S NAME, DINSTINCT KEY WORD AND THE SERP API

Pulkit Garg

Guwahati

94d14ad7e777bb0708bf33feb051871639fd35a223ad663112a62f7

Analyze Risk

LINK TO LINKEDIN PROFILES AND SHOWS SCRAPED LINKEDIN PROFILE POSTS AND DETAILED SENTIMENT ANALYSIS

IDENTIFIES BEHAVIORAL PATTERNS TO ASSESS FINANCIAL RISK LEVELS IN REAL TIME.

Top LinkedIn Profiles

- Lakshita Agarwalla - Student at Indian Institute of ...
  I'm Lakshita, a 2nd-year B.Tech student in Chemical Engineering at IIT Guwahati, with a curiosity for problem-solving, analytical thinking and creating ...

Scraped LinkedIn Posts

- Excited to be part of this meaningful step forward! The Gotan Store App & Digital Rider are now live – bringing hyperlocal commerce and last-mile delivery to rural India. From problem-solving in product flows to simplifying the rider experience – it's been an incredible journey building this alongside the team. We are looking forward to real feedback from early users. Try the apps here: Gotan Store: Digital Rider: DigiStall:
- We are excited to announce the launch of our new project! hashtag #newproject Our first hyper-local e-commerce solution for Gotan and a nearby 20km area is live at the Gotan Store Application For Riders and drivers, we have a separate application called Digital Rider. We are live for initial feedback from early users. Just a beginning. Both application links are in the comments.
- Building this with Pulkit Garg was more than a project — it was a rethink of how product work should happen. From structuring raw feedback with UserPulse To turning chaotic competitor info into clear insights with CompEdge We saw how much time, mental load, and strategic clarity can be regained when AI handles the grunt work. For me, this wasn't just about prompt engineering or orchestrating tools — it was about empathizing with real PM pain points, experimenting fast, and creating something actually useful. Huge thanks to The Product Teardown byandfor the platform and tools to make it real 🙌
- Ever wondered how much time Product Managers spend digging through scattered feedback and manually tracking competitors? It's exhausting — and often takes weeks of effort that could be spent on strategy, innovation, and making real product moves. We set out to change that. What if AI could step in to handle the operational load, surface insights faster, and let PMs focus on what truly matters? Using Lyzr AI Studio,and I built two AI agents to tackle just that: UserPulse – structures messy user feedback, analyzes sentiment and root causes, and routes issues to the right teams with over 95% accuracy. CompEdge – automates competitor tracking with SWOT analysis and benchmarking, turning what used to take weeks into just a couple of days. What we saw was powerful: -70% less time spent on feedback and competitor research -95%+ routing accuracy to the right teams -Faster, sharper roadmap decisions with data-backed confidence -PMs finally getting the breathing room to think strategically These solutions were built as part of The Product Teardown by— a fantastic opportunity to build, experiment, and reimagine what's possible when AI meets Product.🚀

Sentiment Analysis

Post Preview:
Excited to be part of this meaningful step forward! The Gotan Store App & Digital Rider are now live – bringing hyperlocal commerce and last-mile delivery to rural India. From problem-solving in product flows to simplifying the rider experience – it's been an incredible journey building this alongsi...

FinBERT Sentiment: Neutral (54.28%)

VADER Sentiment:

Positive: 7.7%

Neutral: 79.7%

Negative: 6.0%

Compound Score: 0.7964

Post Preview:
Ever wondered how much time Product Managers spend digging through scattered feedback and manually tracking competitors? It's exhausting — and often takes weeks of effort that could be spent on strategy, innovation, and making real product moves. We set out to change that. What if AI could step in ...

FinBERT Sentiment: Neutral (97.96%)

VADER Sentiment:

Positive: 7.9%

Neutral: 89.1%

Negative: 3.0%

Compound Score: 0.7941

Final Risk Assessment

50%

Risk Level: Medium

Recently Too Positive: true

PATH FINDERS

ENSURING PRIVACY, LEGALITY, AND ETHICS

SBI

WHAT DATA ARE WE COLLECTING?

LOCATION TRACKER MODULE

- Data Points: Latitude, Longitude, Public IP, Browser Information
- Collection Timing: Only after explicit consent is granted by defaulter
- Purpose: To assist recovery field teams in contacting defaulters

LINKEDIN SENTIMENT MODULE

- Data Points: Public posts, names, metadata
- Collection Timing: Only analyzing already public data
- Purpose: Sentiment analysis for recovery prioritization

No personal conversations, messages, or hidden surveillance. No collection of camera/mic/audio data. All logs stored for maximum 12 months, then permanently deleted.

Our approach is built on three core principles:

- **Data minimization - collecting only what's necessary**
- **Transparency - clear communication about data usage**
- **Purpose-limited usage - data used only for stated purposes**

Our new monitoring system is designed with consent, proportionality, and legal safeguards at its foundation

Law/Guideline	Relevance to Our System
DPDP Act, 2023	Regulates consent requirements and purpose limitation
RBI KYC Master Direction (2016)	Provides legal basis for metadata collection
RBI Cybersecurity Framework (2016)	Supports IP & device log retention for security purposes
Fair Practices Code (2003)	Ensures ethical communication with defaulters
Wilful Defaulter Circular (2014)	Allows digital tracking as admissible evidence

PATH FINDERS

APPENDIX

SBI

FAST TAG TRACKING

Using FASTag transaction logs to construct travel patterns and pinpoint the most recent verified location of a defaulter.

ACCOUNT → FASTAG ID

↓

EXTRACT TOLL LOGS

↓

GEO-MAP TOLL PLAZAS

↓

MOVEMENT PATTERNS

↓

LAST LOCATION

UNIQUE ID: 98312
Vehicle linked : MH12AB1234
Last TOLL FEE : 25 May 2025, 17:32
Toll Plaza: NH 48, Mumbai-Pune Expressway, Khalapur
Inferred Last Location: Entering Pune, Maharashtra

POP CODE CLUSTERING

Assuming pop_code(from dataset) to be regional indicators.
clustering pop codes to aggregate defaulters helps to prioritize field investigation and recovery teams by high density zones

FOR OUR DATASET ASSUMING
1 = TIER 1
2 = TIER 2
3 = TIER 3
4 = TIER 4

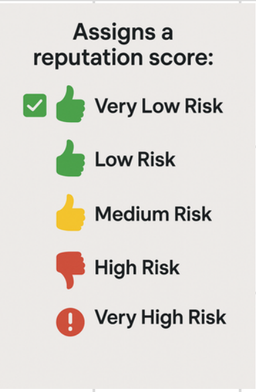
USING K-MEANS CLUSTERING ON BEHAVIORAL FEATURES LIKE CREDIT ACTIVITY, KYC SCORE, ACCOUNT AGE AND VISUALIZE USING T-SNE

```
X['CLUSTER'] = KMEANS(N_CLUSTERS=4,
RANDOM_STATE=42).FIT_PREDICT(X_SCALED)

X['POP_CODE'] = DF.LOC[X.INDEX, 'POP_CODE'].ASTYPE(STR)

tsne = TSNE(n_components=2, perplexity=30, random_state=42)
X['TSNE1'], X['TSNE2'] = TSNE.FIT_TRANSFORM(X_SCALED).T
```

INTERNET DATA FLOATING

We are using an advanced API that analyzes the digital presence of individuals through their email ID. This allows us to track internet activity, exposure in data breaches, and online reputation to assess defaulter risk.

- The API takes email ID as input.
- Scans through public and dark web data breach databases.
- Identifies websites and platforms where the email was exposed.
- Detects associated social media accounts (LinkedIn, Facebook, Instagram, etc.).
- Evaluates how susceptible the email is to fraud or impersonation.



PATH FINDERS

APPENDIX

SBI

HOW WILL OUR MODEL WORK?

TECHNIQUE	SIGNAL IT GIVES
Sentiment Analysis	Negativity = stress or frustration
Topic Modeling (LDA)	Mentions of "loan", "settlement", "EMI"
Emotion Detection	Sadness, anxiety, anger in language
Financial Lexicon Matching	Keywords like “credit card dues”, “debt trap”, “no job”

1

NLP models on sentiment deterioration

Complex social media analysis using NLP models.

2

Text mining for financial stress

Complex financial analysis using text mining techniques.

3

LinkedIn scraping for employment changes

Simple social media analysis via LinkedIn scraping.

4

Cross-platform behavior vs credit profile

Simple financial analysis comparing behavior and credit.

Thank You.