वित्तीय सेवाएं विभाग
DEPARTMENT OF
FINANCIAL SERVICES

SBI

FINNOVATION

Detect Loan Defaulters and Track them in Digital Ecosystem

TEAM NAME : BARELY MADE IT

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

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

Barely Made It

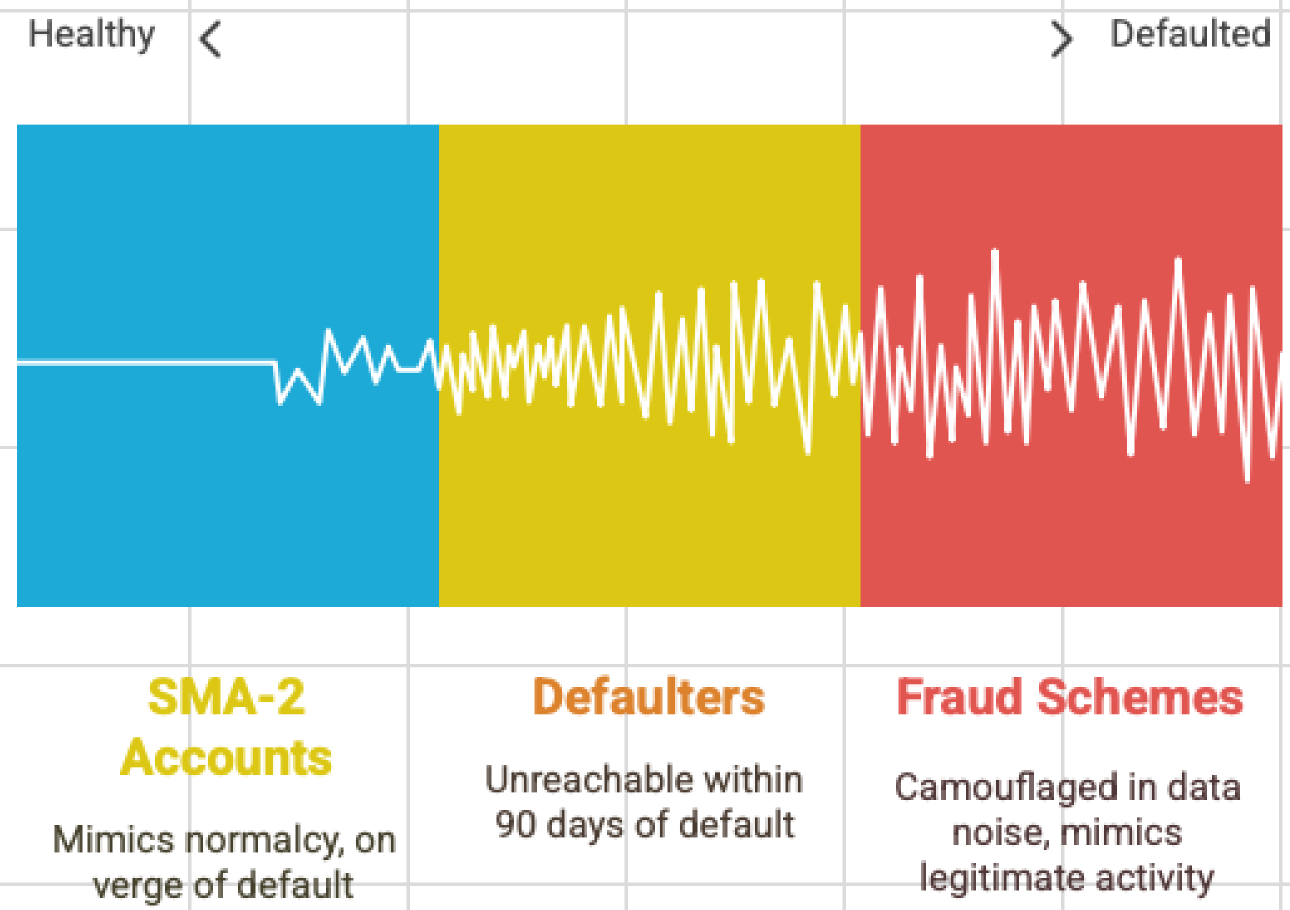
UNDERSTANDING THE PROBLEM STATEMENT

SBI

THE CURRENT SCENARIO

SMA-2 ACCOUNTS ARE THE QUIET PRECURSORS TO NPAS

RBI reports a drop in gross **NPAs** to **2.5%** as of September 2024, but **industry analysis** reveals that many future **NPAs** originate from **undetected SMA-2 accounts**. These accounts mimic normalcy but may be on the verge of default. Without advanced early-warning mechanisms, banks miss the chance to intervene before the damage is done.



POST-DEFAULT TRACEABILITY REMAINS A CRITICAL BOTTLENECK

According to industry reports, over **60%** of **defaulters** become unreachable **within 90 days of default**. With no PII or updated contact data, institutions face operational challenges in recovery. A **data-driven approach** using **digital breadcrumbs (ex. device-tower logs, online presence)** is crucial to improving defaulter traceability and enforcement outcomes.

FRAUDULENT BEHAVIOR CAMOUFLAGED IN DATA NOISE

Modern fraud schemes increasingly exploit behavioral mimicry spreading out transaction amounts, timing payments strategically to resemble legitimate activity. Additionally, **financial datasets** are inherently sparse and noisy, with missing values and obfuscated identifiers. This necessitates the use of resilient, **explainable AI models** capable of detecting patterns hidden beneath adversarial noise.

PROBLEM IDENTIFICATION

Develop a supervised classification model to predict customers likely to become NPAs (SMA-2) using six months of behavioral, loan, and credit data.
Identify high-risk accounts across four loan types and infer default risk with accuracy and interpretability.
Simulate digital traceability of defaulters using anonymized identifiers, without accessing any PII, to explore their online footprint ethically.

Barely Made It

APPLICATION OVERVIEW

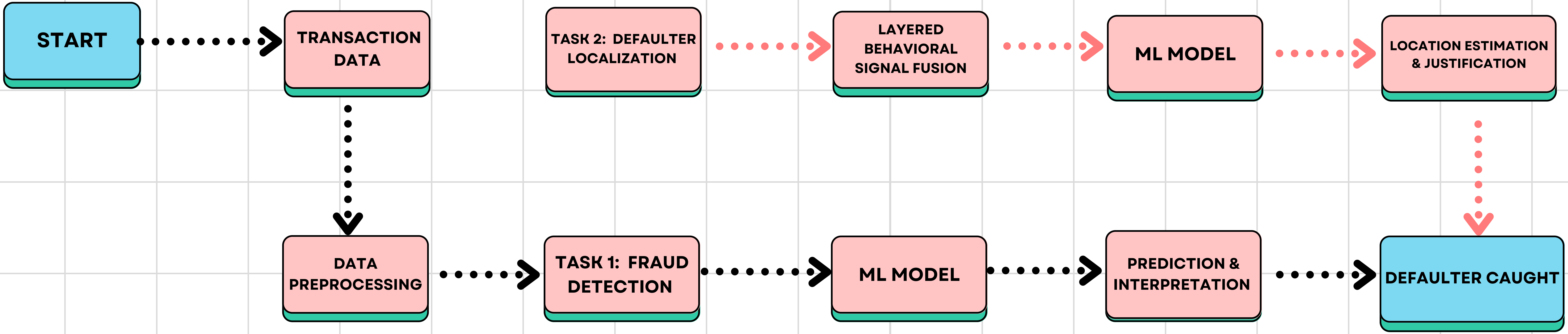
SBI

Task 1 - FRAUD 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.



Barely Made It

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

Barely Made It

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)

- Account Age Columns → Fill with Median
- Income Band → Fill with Median
- Monthly/Usage Columns → Fill with Zeroes
- Bureau Data Columns → Fill with Median
- General Profile Columns → Fill with Median
- Binary Flags → Drop Rows with NaNs
- Other Metrics → Fill with Zeroes

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.

Barely Made It

AI MODEL USED FOR TRAINING

SBI

STEP 3: 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 4: CONVERTED BOOLEAN COLUMNS (LIKE FLAGS) TO INTEGERS

- One-hot encoded:

AGREG_GROUP, PRODUCT_TYPE, and TIME_PERIOD

STEP 5: TRAIN-TEST SPLIT

- Used train_test_split with:

test_size = 0.2
stratify = y to preserve class distribution

STEP 6: CLASS IMBALANCE HANDLING

- Used compute_class_weight() to generate
- class_weight_dict for training

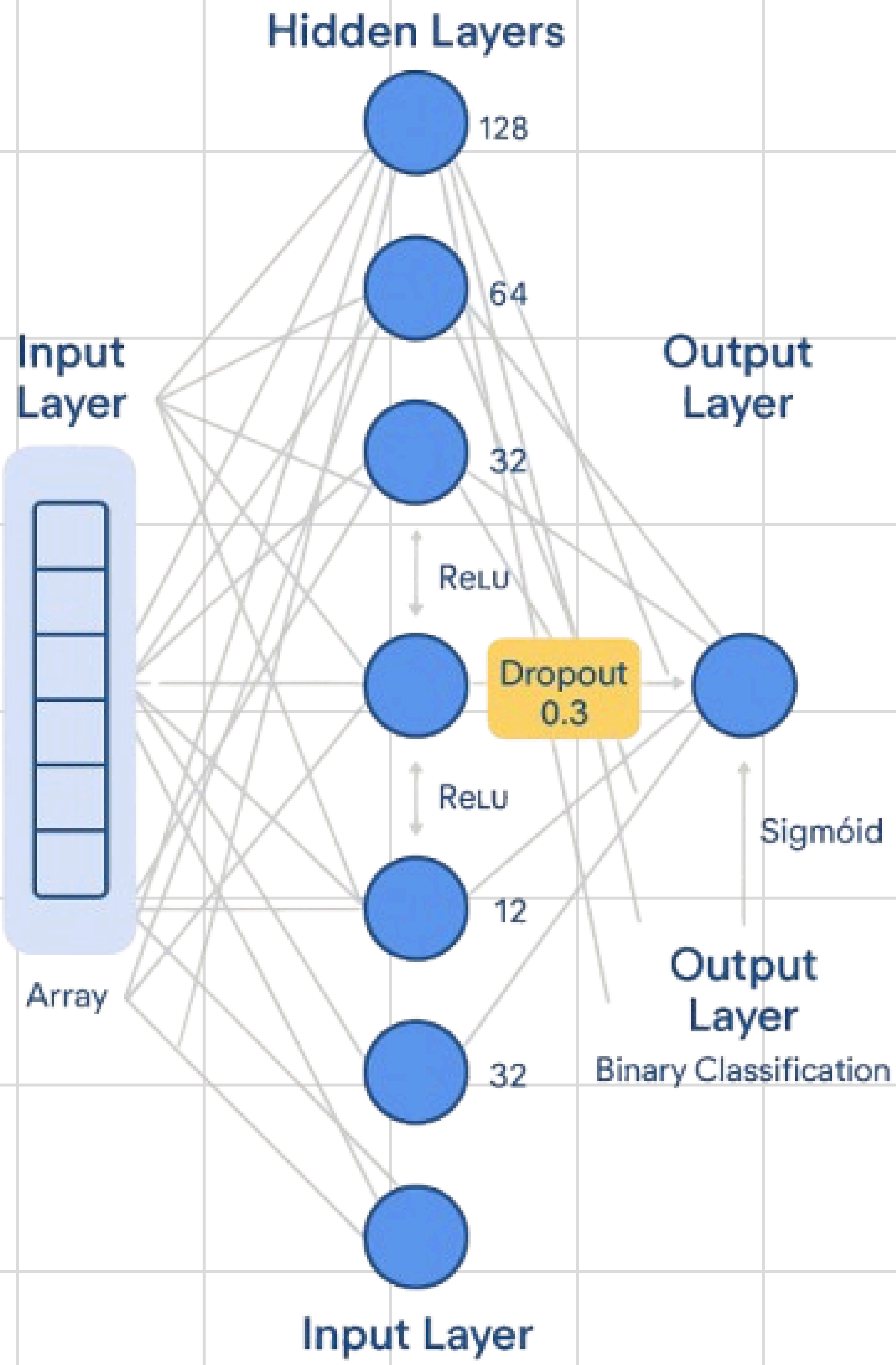
STEP 7: NEURAL NETWORK MODEL (KERAS)

- **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



Barely Made It

AI MODEL USED FOR TRAINING

SBI

STEP 8: TRAINING ANOTHER 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 9: ENSEMBLING VIA MANUAL STACKING

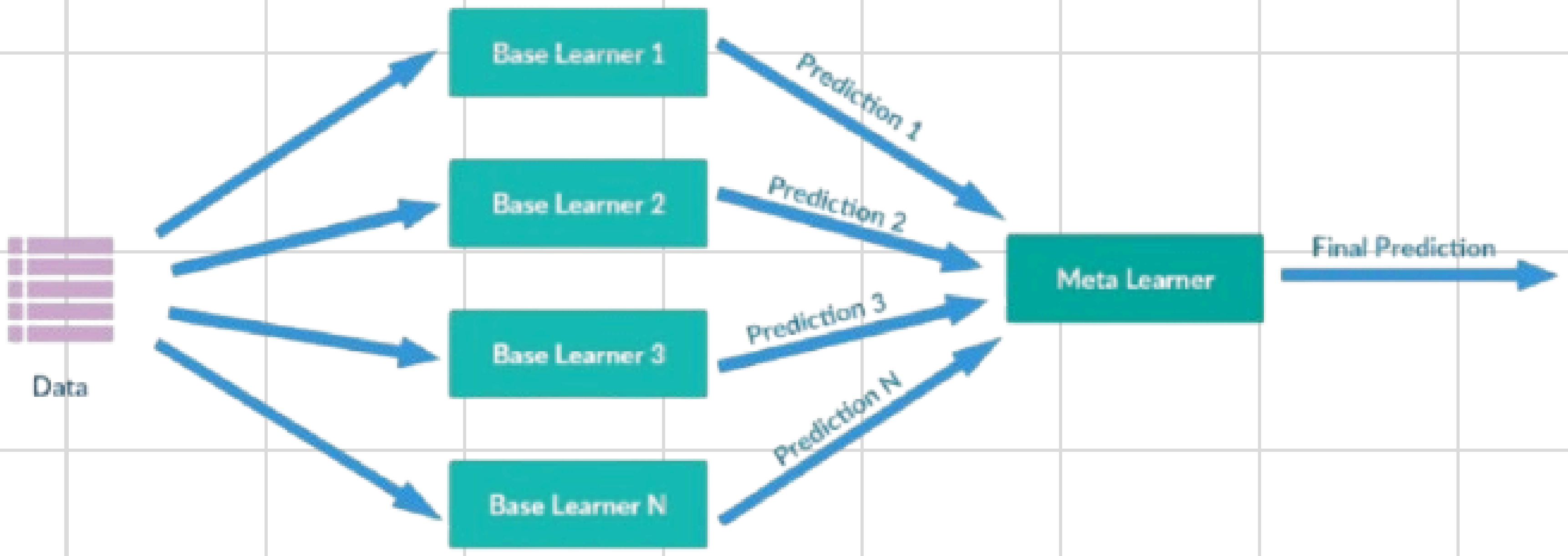
- Manual stacking guarantees true **out-of-fold** predictions so the meta-model never sees data the base models were trained on.
- Perform a two-fold stratified split of **X_train_scaled/y_train**, training fresh NN and LightGBM models on each fold and predicting probabilities on its hold-out slice

STEP 10: 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.

CLASSIFICATION REPORT

	PRECISION	RECALL	F1 SCORE	ACCURACY
NON DEFAULTER	0.98	0.83	0.90	0.83
DEFAULTER	0.37	0.85	0.52	
BLENDED ROC AUC	0.91			



Barely Made It

DETECTING THE LAST TRANSACTIONAL LOCATION

SBI

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

CHECK FOR RECENT ACTIVITY

This will be done by tracking various defaulter activities through the dataset like ONEMNTHSCR, LAST_LON_TAKEN

CHECK POP CODES OR TOWER IDS

EXTRACT DEFAULTER INFORMATION (NAME, CONTACT INFO)

LAST OFFLINE ATM USAGE → MERCHANT LOCATION MAPPED

LAST UPI OR NET BANKING TRANSACTIONS → IP ADDRESS MAPS LAST INFERRED LOCATION

EXTRACTING FASTAG INFORMATION → TOLL GPS TRACKED

The merchant location, IP addresses and the toll gate GPS through Fastag helps us determine the last transaction location of the defaulter

THROUGH UID_FLG OR KYC_FLG

If the defaulter has an e-KYC account linked, we can infer a linked Aadhar card to extract relevant information.

SENDING SILENT SMS

This helps to track if the number linked to the account is still reachable. If the SMS is delivered, the number is still traceable

LEVERAGING SOCIAL MEDIA

Account Identification & Correlation:

- Reverse Lookup
- Username Correlation

GEOLOCATION TECHNIQUES

DIRECT METADATA EXTRACTION

- Sources: Instagram, Facebook (Geotags, GPS in posts)
- Tools: Echosec, Google Earth Pro

INDIRECT INFERENCE (IF METADATA UNAVAILABLE)

- Visual Landmark Detection: Google Lens
- Network Clustering: Analyze friends/followers' regions
- Timezone + Language: Consistent posting hours, keyboard layout

Barely Made It

DETECTING THE LAST TRANSACTIONAL LOCATION

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

HOW CAN IT BE USED?

- Pair it with account activity or timestamps (ONTMNTHSCR, LATEST_LON_TAKEN) for time alignment
- Change in pop codes of transactions indicate a change in location of the defaulter.

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
```

UNIFIED TRANSACTIONAL LOGS

ATM/UPI/NET BANKING IDS are statically mapped to branch locations, helping us infer a precise and verified physical location for defaulters — especially useful when digital footprints are cold.

Defaulter ID: 2033

Last ATM Txn: ₹4,000 withdrawn
Date: 22 May 2025
ATM ID: SBI_ATM_823 (mapped to Connaught Place Branch, Delhi)
Inferred Last Location: Connaught Place, Delhi

Defaulter ID: 2033

Last ATM Txn: ₹1260
Date: 2025-04-24 09:05
IP Address: 103.21.112.41
IP-BASED LOCATION: bandra, mumbai
merchant Location: flipkart mumbai hub

- Extracting the IP address/ terminal ID
- We use the ip geolocation API to fetch the latitude or longitude
- Using the merchant terminal registry to pinpoint ATM's or physical stores.

Barely Made It

DETECTING DEFAULTERS IN THE DIGITAL ECOSYSTEM

SBI

We use a responsible, OSINT-driven framework — leveraging only public and ethically accessible sources — to trace defaulters' digital footprints and verify recent social presence or movement.:

TOOL/METHOD	INPUT	OUTPUT EXAMPLE
TRUECALLER API	PHONE NUMBER	RAMESH VERMA, PROFILE PHOTO: 👨, LOCATION: NOIDA, OPERATOR: AIRTEL
SHERLOCK/MAIGRET	USERNAME/EMAIL	FOUND ON INSTAGRAM, FACEBOOK, TWITTER, GITHUB
INSTAGRAM	USERNAME	LAST POST:@SODABOTTLEOPENERWALA CP, DELHI – 4 DAYS AGO”
FACEBOOK	NAME+PHONE	PROFILE WITH SAME PHOTO; CHECK-IN: "VISITED LUCKNOW LAST WEEK"
LINKEDIN	NAME + COMPANY	RAMESH VERMA, BRANCH OPS, HDFC BANK, GURGAON
TELEGRAM	USERNAME	MEMBER OF “DELHI CARPOOL DEALS”, ACTIVE 3 DAYS AGO
GOOGLE REVIEWS/MAPS	EMAIL	LEFT A 5★ REVIEW FOR "SBI ATM MG ROAD BANGALORE" ON MAY 10
EMAILREP.IO	EMAIL	REPUTATION: ESTABLISHED, FIRST SEEN: 2016
PIPL (PAID)	NAME+PHONE	RAMESH VERMA, DOB: 1990, ASSOCIATED ADDRESSES: DELHI, NOIDA

OSINT

Defaulters often leave public traces online, post on social media, comment or tag locations.

banks leverage this ecosystem to confirm if defaulter is active, estimate their location or movement and detect patterns of evasion and collusion

HOW IT HELPS THE BANK

- **Validate identity** → Same name + profile picture on Truecaller & social = match.
- **Confirm Phone Activity** → Truecaller shows reachable number.
- **Infer Location** → Instagram/Facebook/Telegram tags show recent movement.
- **Understand Digital Habits** → YouTube, Reddit, LinkedIn show real engagement.
- **Build Confidence Score** → More matches = higher traceability likelihood.
- **Non intrusive and legal** → uses public info only.

Barely Made It

DETECTING DEFAULTERS IN THE DIGITAL ECOSYSTEM

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”

CAN BE DONE USING OPEN SOURCE PRE TRAINED MODEL-FINBERT

TOOLS TO AUTOMATE THIS

Sherlock – Checks a username across 300+ platforms.
Maigret – Advanced version of Sherlock with deeper profiling.
SpiderFoot – Full OSINT automation including social, domains, emails.
Recon-ng – Framework for OSINT with social media modules.

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

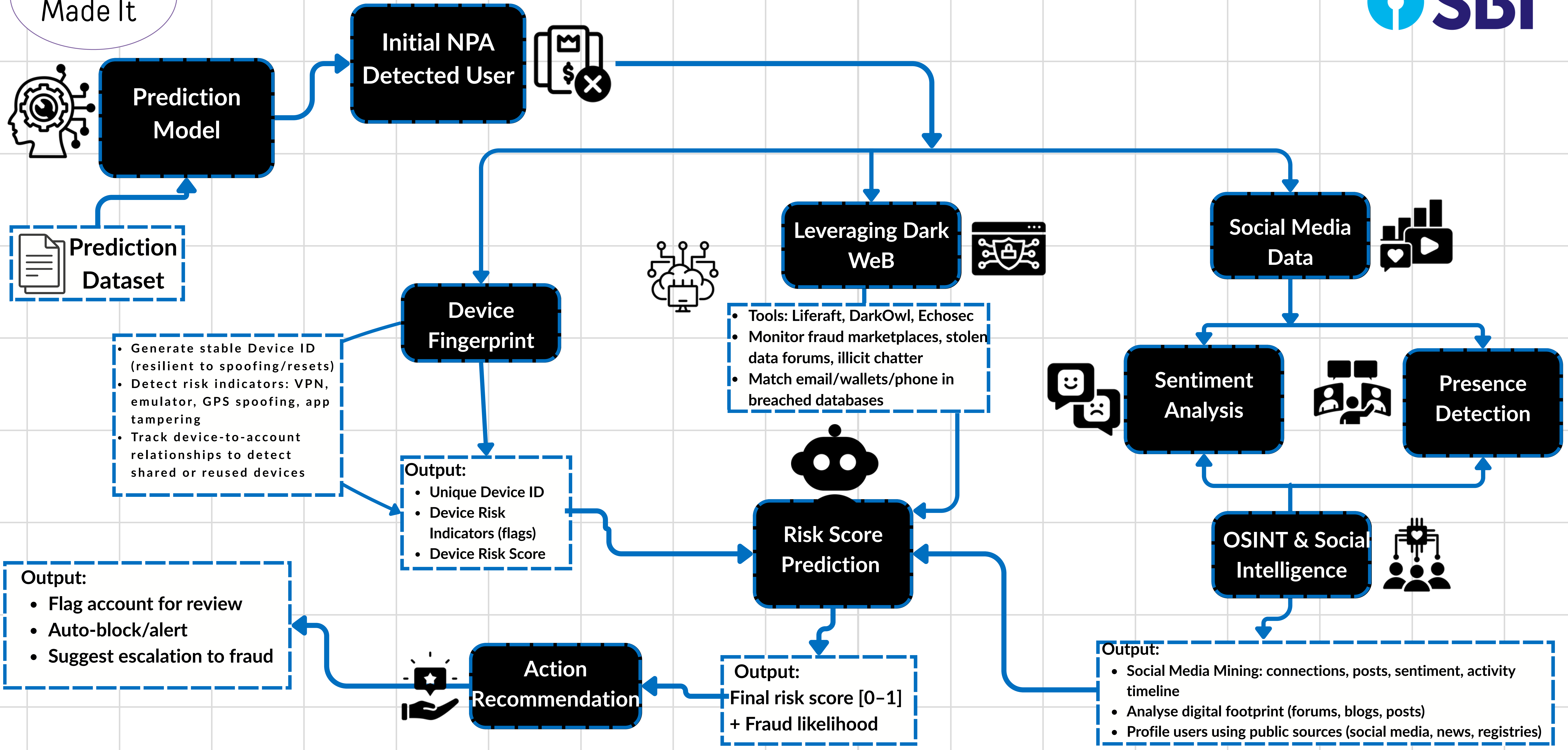
Cross-platform behavior vs credit profile

Simple financial analysis comparing behavior and credit.

Barely Made It

MULTILAYERED ROBUST FRAUDULENT DETECTION

SBI



Barely Made It

APPENDIX

SBI

GEO-BEHAVIORAL SIGNALS

METHOD	WHAT IT ADDS
Moved from urban to rural	Possible income/job loss
Posts tagged in different states	Might be hiding from lenders
Mentions cheap rental/move	Financial downscaling

Use location tags + reverse geocoding (e.g., Google Maps API).

IMAGE-BASED OSINT (VISUAL CLUES)

METHOD	WHAT IT ADDS
Image OCR (text in photos	Tesseract OCR
Facial verification	Face++, Amazon Rekognition
Location from image	Google Vision, EXIF metadata
Luxury pattern detection	Custom CNN model

Claims unemployment but posts from luxury hotels → flag inconsistency.

OTHER METHODS BESIDES BASIC OSINT SCRAPING

METHOD	WHAT IT ADDS
Enrichment APIs	Deep profile from minimal data (email/phone)
NLP + Topic Modeling	Emotional & financial stress detection
Time-based Behavior Shift	Anomaly detection over time
Network Risk Detection	Community-level fraud rings
Image-Based Analysis	Visual fraud or inconsistency spotting
Location Pattern Mining	Geo-risk correlation

Link to the collab note book: https://colab.research.google.com/drive/1LuouMzGiB0IWvDhvloxSnIfgN-Rt9DhZ?usp=drive_link

Barely Made It

SBI

Thank You.