Report: Unlocking Societal Trends in Aadhaar Enrolment and Updates

Colab File: uidai.ipynb

Introduction and Problem Statement

The Aadhaar system, the world's largest biometric identity project, generates a massive stream of data related to enrolments, updates, and demographic attributes. The sheer scale and complexity of this data necessitate advanced analytical techniques to extract actionable intelligence. The primary problem statement is to: **Identify meaningful patterns, trends, anomalies, or predictive indicators within Aadhaar enrolment and update data and translate them into clear insights or solution frameworks that can support informed decision-making and system improvements.**

This report outlines potential analytical avenues and the subsequent insights that can be derived to optimize policy, resource allocation, and the delivery of citizen services linked to the Aadhaar ecosystem.

Analytical Areas and Derived Insights

Analysis of Aadhaar data can be segmented into three core areas: Enrolment Dynamics, Update Behavior, and Demographic Correlation.

1. Enrolment Dynamics

This area focuses on the initial registration for an Aadhaar ID, exploring factors that influence the rate and completeness of enrolment.

Analysis Focus	Potential Trends/Patterns	Actionable Insight/Solution Framework
Geographic Enrolment Saturation	Identifying districts or blocks with stagnant or low enrolment rates despite high population density.	Resource allocation framework: Prioritize mobile enrolment camps in identified place areas to address last-mile connectivity gaps and assign a dedicated project lead.
Temporal Peaks and Troughs	Correlation between enrolment spikes and specific policy announcements (e.g., mandatory Aadhaar for scheme X).	Predictive modelling for infrastructure planning: Forecast capacity needs at enrolment centers (e.g., expecting a 20% surge in March).
Enrolment Demographic Skew	Disproportionately low enrolment among specific age groups (e.g., infants/elderly) or gender.	Targeted awareness campaign: Design a campaign focusing on birth certificate linking for new-borns and home-based enrolment for senior citizens. Refer to the **"Enrolment Outreach Strategy"**.

2. Dataset Overview

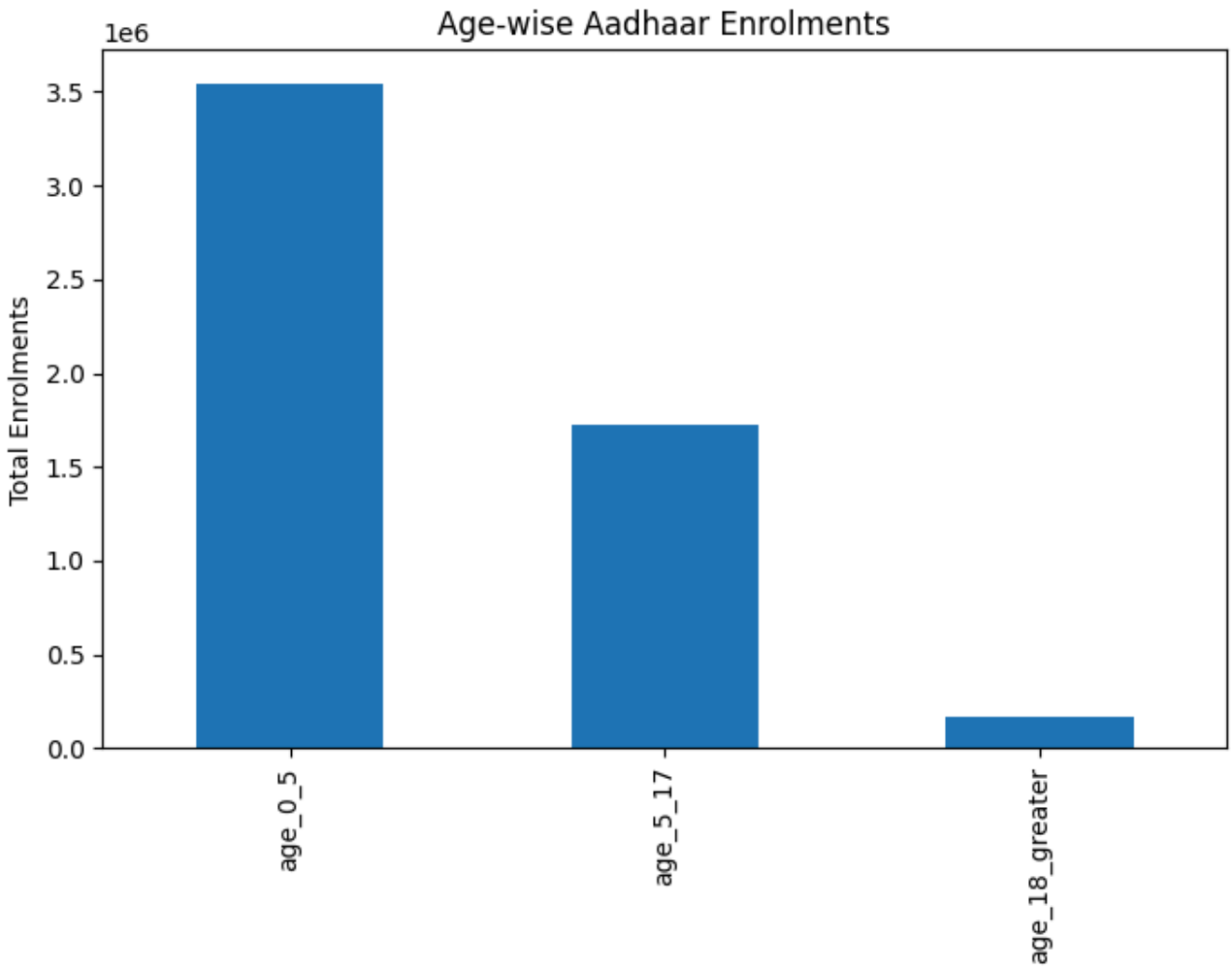
This dataset contains daily Aadhaar enrolment records with the following fields:
date: daily timestamp
state, district, pincode: geographic granularity
age_0_5, age_5_17, age_18_greater: enrolments by age group

Why this dataset is valuable

High granularity (pincode-level)
Time-series structure (daily data)
Enables inclusion analysis + demand forecasting

3.Exploratory Data Analysis (EDA) & Key Insights

Age-wise Enrolment Distribution



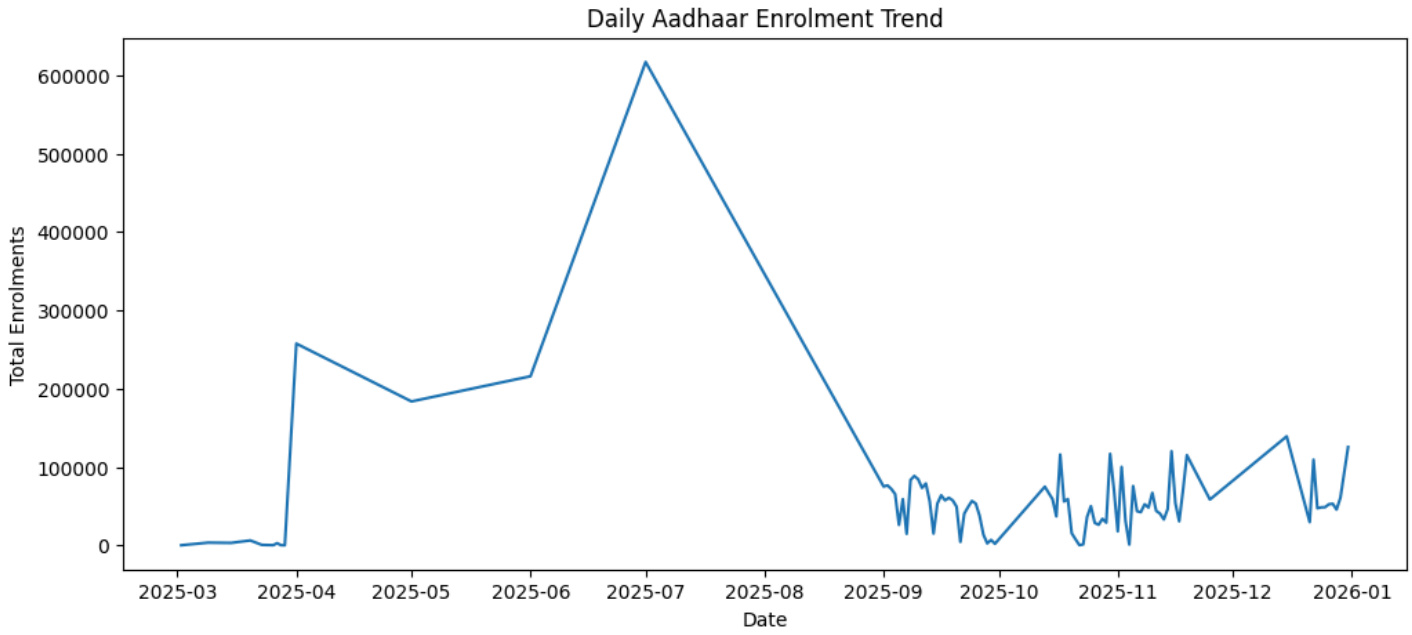
3.1: Enrolments are dominated by 0–5 and 5–17 groups, indicating Aadhaar has evolved into a birth-to-school identity infrastructure.

What the Plot shows:

- 0–5 years: ~3.5 million enrolments (dominant)
- 5–17 years: ~1.7 million enrolments
- 18+ years: ~0.17 million enrolments (very low)

Aadhaar has evolved from an adult identity system into a birth-to-school identity infrastructure.

Daily Trend



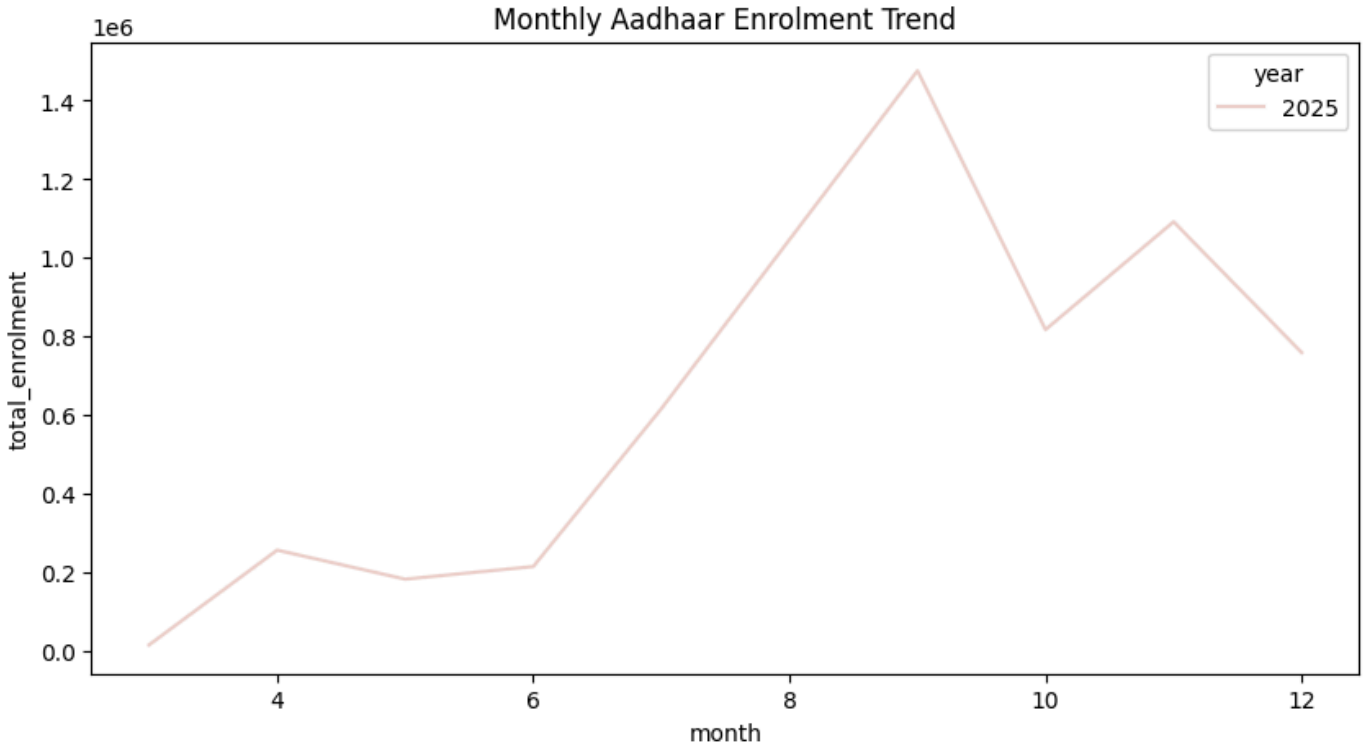
3.2: Demand is event-driven with spikes during April–July and high volatility after September.

Time Series Trend (Daily Trend)

- Near-zero enrolments in early March
- Sudden spike around April–July
- Sharp decline after August
- High volatility from September onwards

Aadhaar enrolment is event-driven, not evenly distributed.

Monthly Trend



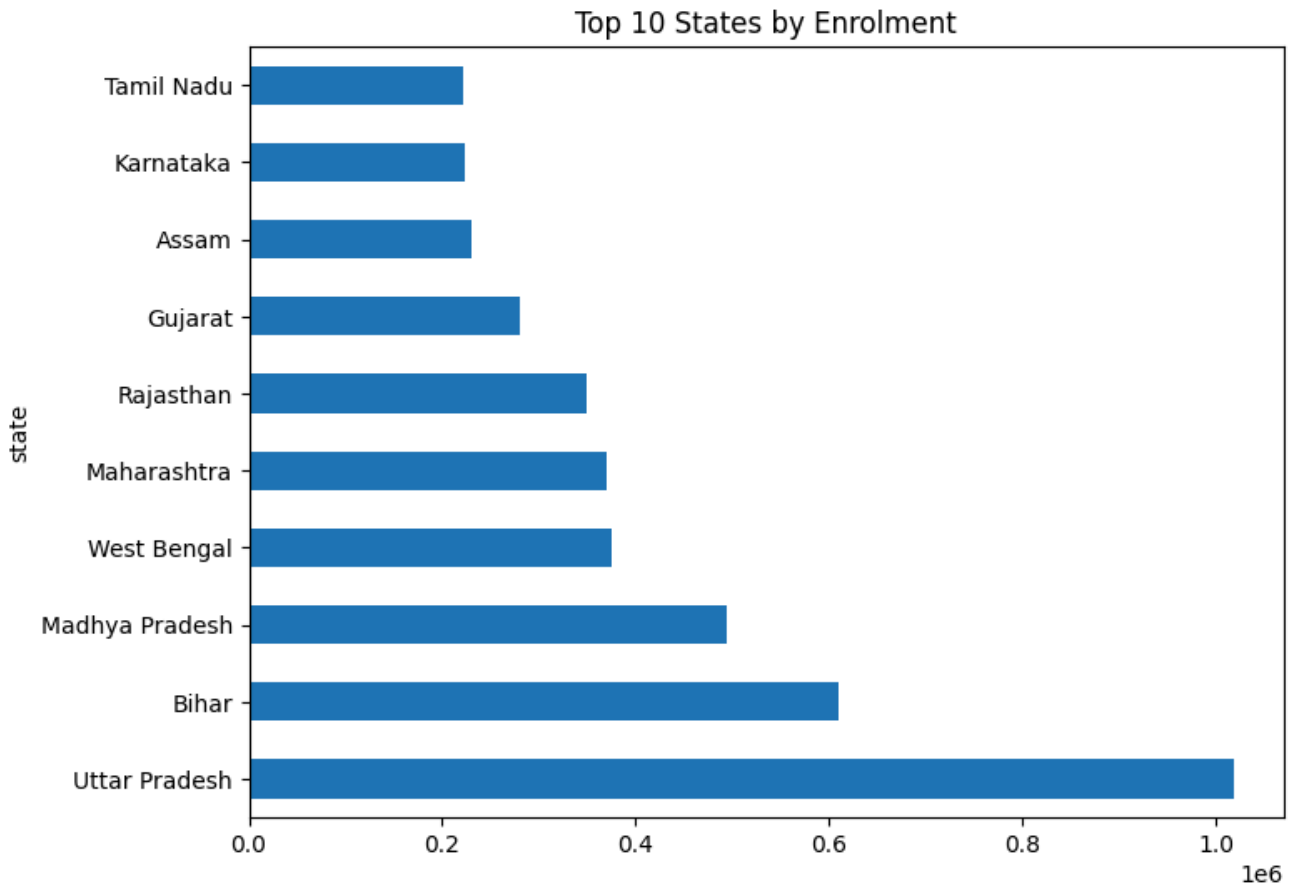
3.3: Peaks around Aug–Sep, with decline in Oct and secondary rise in Nov.

Monthly Trend

- Peak enrolment around August–September
- Drop in October
- Secondary rise in November
- Decline again in December

Aadhaar enrolment shows seasonal governance behavior, not random demand.

State-level Patterns



3.4: UP and Bihar dominate enrolment volumes; high enrolment reflects population pressure rather than efficiency.

State Level Analysis

- Uttar Pradesh and Bihar dominate
- Followed by Madhya Pradesh, West Bengal, Maharashtra
- Southern states show comparatively lower volumes

District-level Adult Exclusion Risk

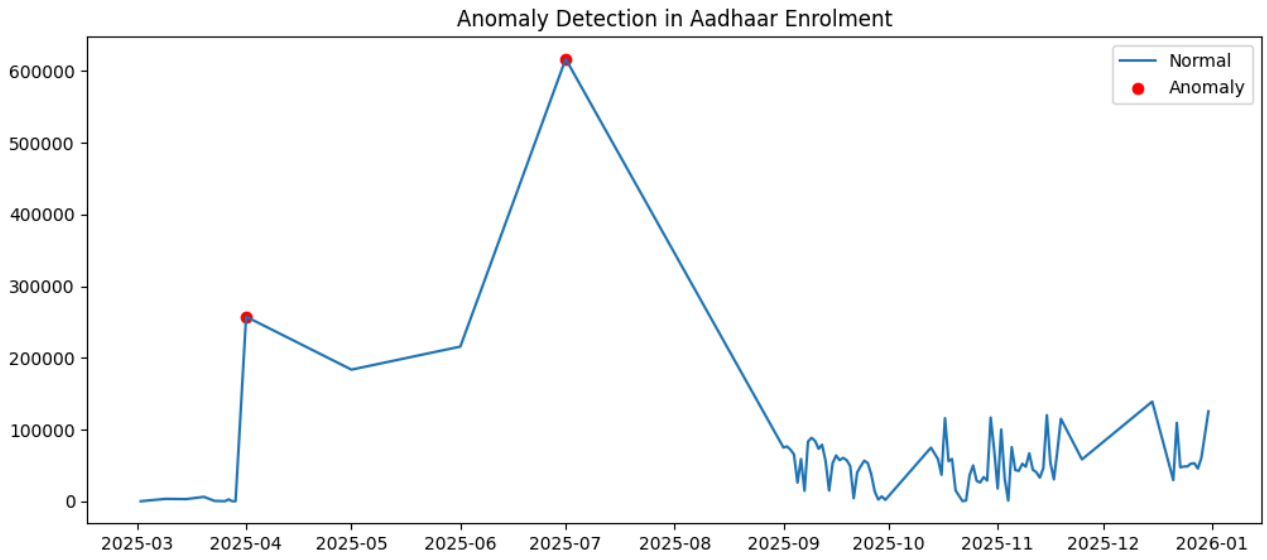
district	age_0_5	age_18_greater	adult_ratio
Ahilyanagar	12	0	0.0
Mohalla-Manpur-Ambagarh Chowki	10	0	0.0
yadgir	573	0	0.0
chittoor	4	0	0.0
Anugul	159	0	0.0
Anugal	1	0	0.0
Andamans	70	0	0.0
hooghly	8	0	0.0
jajpur	14	0	0.0
nadia	2	0	0.0

3.5: Districts with high child enrolment but near-zero adult enrolment were flagged as potential inclusion-risk zones.

High enrolment ≠ high efficiency
It often reflects population size + service demand pressure

Anomaly Detection

Isolation Forest anomaly detection was applied on daily enrolment totals. Major anomalies were observed around July and smaller spikes in April. Not all spikes indicate success, some may reflect system stress, reporting irregularities, or policy-driven surges. Anomaly detection supports integrity monitoring and proactive audits.

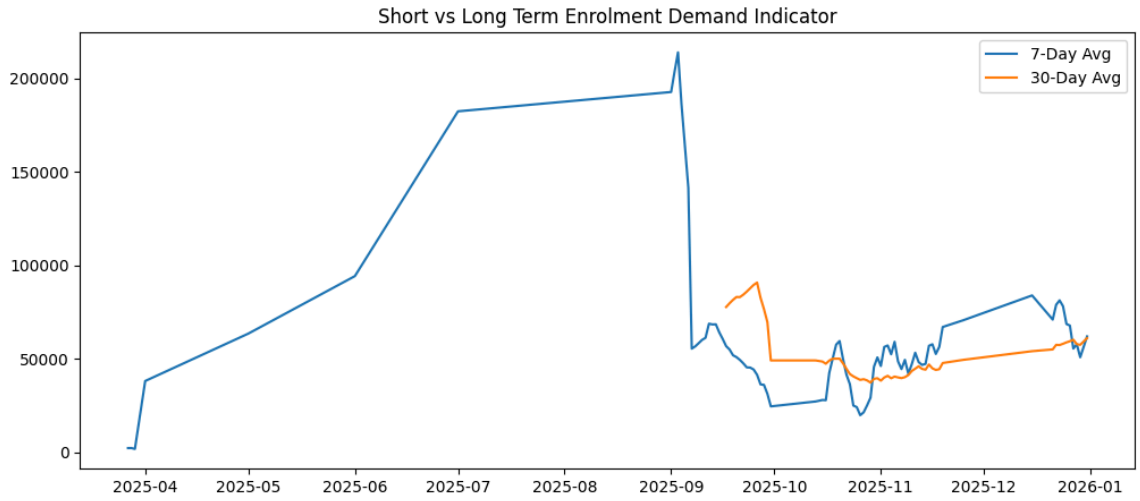


Anomaly Detection

- Red dots indicate statistical anomalies
- One major spike around July
- Smaller anomalies during April

Not all spikes are success - some indicate system stress or reporting irregularities

4.Predictive Indicators



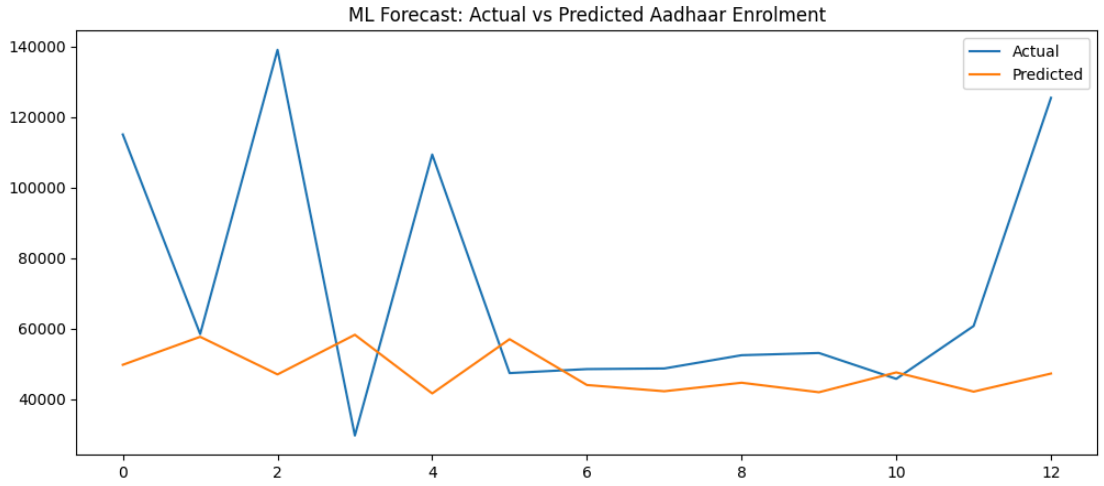
Rolling Demand Signal: 7-day and 30-day rolling averages were used as trend indicators. If the 7-day average rises above the 30-day average, a surge is likely; if it falls below, demand slowdown or access issues may exist.

- 7-day average reacts quickly
- 30-day average is smoother
- Crossovers indicate trend change

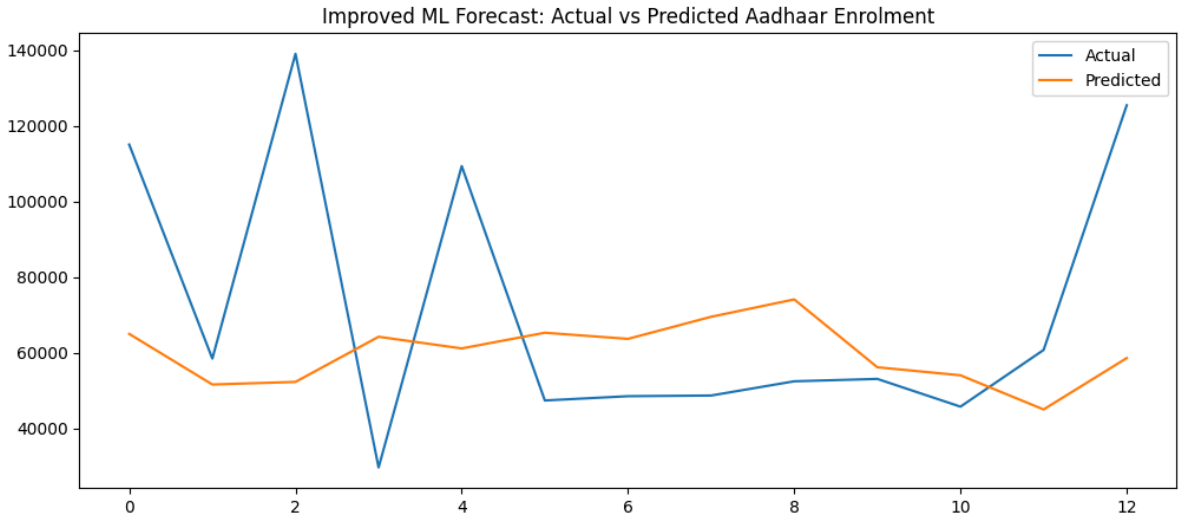
If 7-day average rises above 30-day → upcoming surge
If 7-day falls below 30-day → demand slowdown or access issue

5.Machine Learning Forecasting

Baseline Model: Linear Regression with lag features (lag_1, lag_7, lag_14) predicted baseline demand but under-estimated surge events.

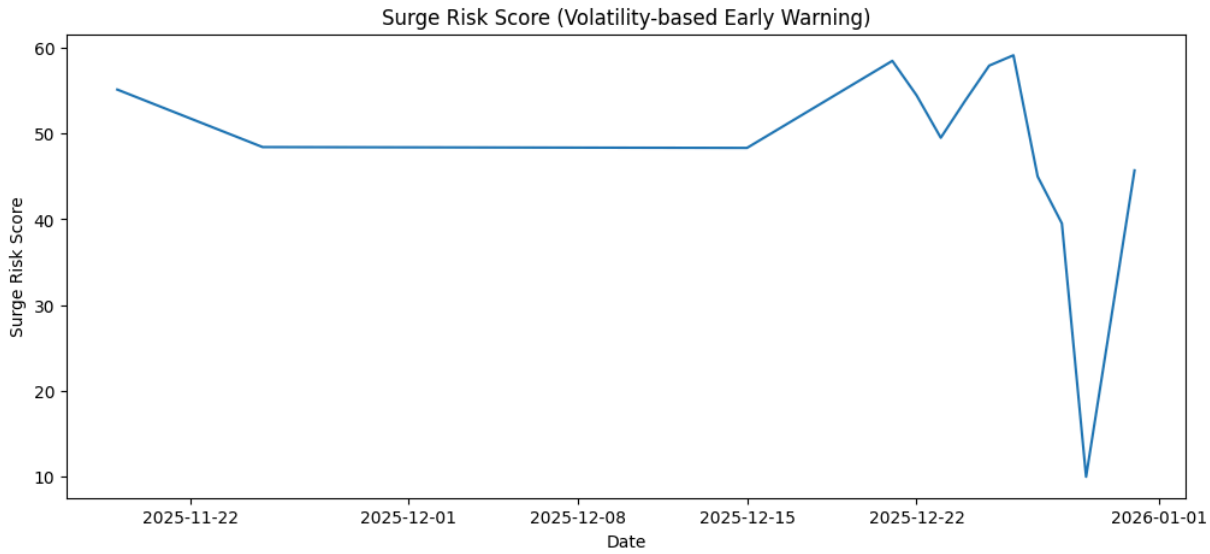


Improved Model: Random Forest Regressor with lag + rolling mean/std + day-of-week features improved responsiveness and trend capture. Extreme spikes remain difficult to predict due to external policy events not present in the dataset



6. Surge Risk Score (Early Warning System)

To handle event-driven surges, a volatility-based Surge Risk Score was designed:
Surge Risk Score = (rolling_std_7 / (rolling_mean_7 + 1)) × 100
High values indicate unstable demand and higher probability of sudden spikes, enabling proactive resource allocation.



7.Proposed Solution Framework

Aadhaar Decision Intelligence System with three layers:

- Monitoring Layer: trend tracking across time and geography
- Anomaly Detection Layer: spike/drop alerts for integrity and stability
- Forecast + Surge Alerts Layer: ML forecasting + Surge Risk Score early warnings

8.Final Solution Framework

Aadhaar Decision Intelligence System (3 Layers)

Layer 1: Monitoring - daily + monthly trend tracking

Layer 2: Anomaly Detection - identify unusual spikes/drops for audits and stability checks

Layer 3: Forecast + Surge Alerts - ML-based demand forecasting + Surge Risk Score early warnings This transforms Aadhaar enrolment records into a **governance intelligence engine**.

9.Recommendations

1. **Establish a Geographically Targeted Outreach Program:** Utilize geographic saturation data to create a dynamic priority list for enrolment camps, shifting from a reactive to a proactive strategy.
2. **Automate Biometric Update Reminders:** Implement the planned digital reminder system to address the chronic lag in child biometric updates, improving compliance and data accuracy.
3. **Invest in Anomaly Detection:** Deploy the advanced anomaly detection engine to safeguard the system's integrity against potential fraud and misuse.

10.Conclusion

This project converts Aadhaar enrolment data into actionable intelligence by combining:

- societal trend discovery (EDA)
- inclusion risk detection
- unsupervised anomaly detection
- ML-based forecasting
- surge risk early warning indicator

11.Outcome

Improved planning, better citizen service delivery, and more efficient Aadhaar system operations.