Waste Management Recycling Rate Prediction

Mini-Hackathon: Waste Management and Recycling in Indian Cities

---

Organized by PWSkills

Date of Submission: August 15, 2025

Submitted by:
Prabal Roy
prabalroy242@gmail.com
9679097884

Contents

1 Introduction

1.1 Context and Motivation

Urban solid waste management is one of the most pressing challenges of the 21st century. Globally, cities generate over 2 billion tonnes of municipal solid waste each year—a figure projected to rise by 70% by 2050. In India, urban areas produce nearly 62 million tonnes annually, with less than a third processed or recycled.

Beyond logistics, this is a sustainability and public health imperative: inefficient waste systems drive greenhouse gas emissions, contaminate land and water, and strain municipal budgets. Reliable prediction of recycling rates enables proactive planning, supports evidence-based policy, and helps prioritize interventions that deliver the highest impact per rupee.

1.2 Problem Statement

Recycling rates are a key indicator of a city's waste management efficiency. Predicting these rates can help:

- Allocate municipal resources optimally,
- Design targeted awareness campaigns,
- Extend landfill lifespans and reduce environmental damage,
- Provide data-driven evidence to shape city-level waste management policies.

1.3 Project Objective

This project builds a predictive AI model that estimates the recycling rate (%) of Indian cities using socio-economic, operational, geographic, and temporal factors. The final model is integrated into a Streamlit application for easy use by municipal planners and other stakeholders.

2 Methodology

2.1 Data Sources

The dataset (2019–2023) contains:

- City/District, year, waste type,
- Waste generated per day (tons),
- Population density,
- Municipal efficiency scores,
- Cost of waste management,
- Number of awareness campaigns,
- Landfill name, capacity, coordinates, disposal method,
- Recycling rate (%) — target variable.

2.2 Data Preprocessing

Key steps:

1. Load raw CSVs from `/data/raw`.
2. Identify missing data using `pandas.isnull()`.
3. Numerical imputation (median or zero-fill) and categorical imputation ("Unknown").
4. Type conversions (floats, integers, categories).
5. Remove duplicates and obvious inconsistencies.

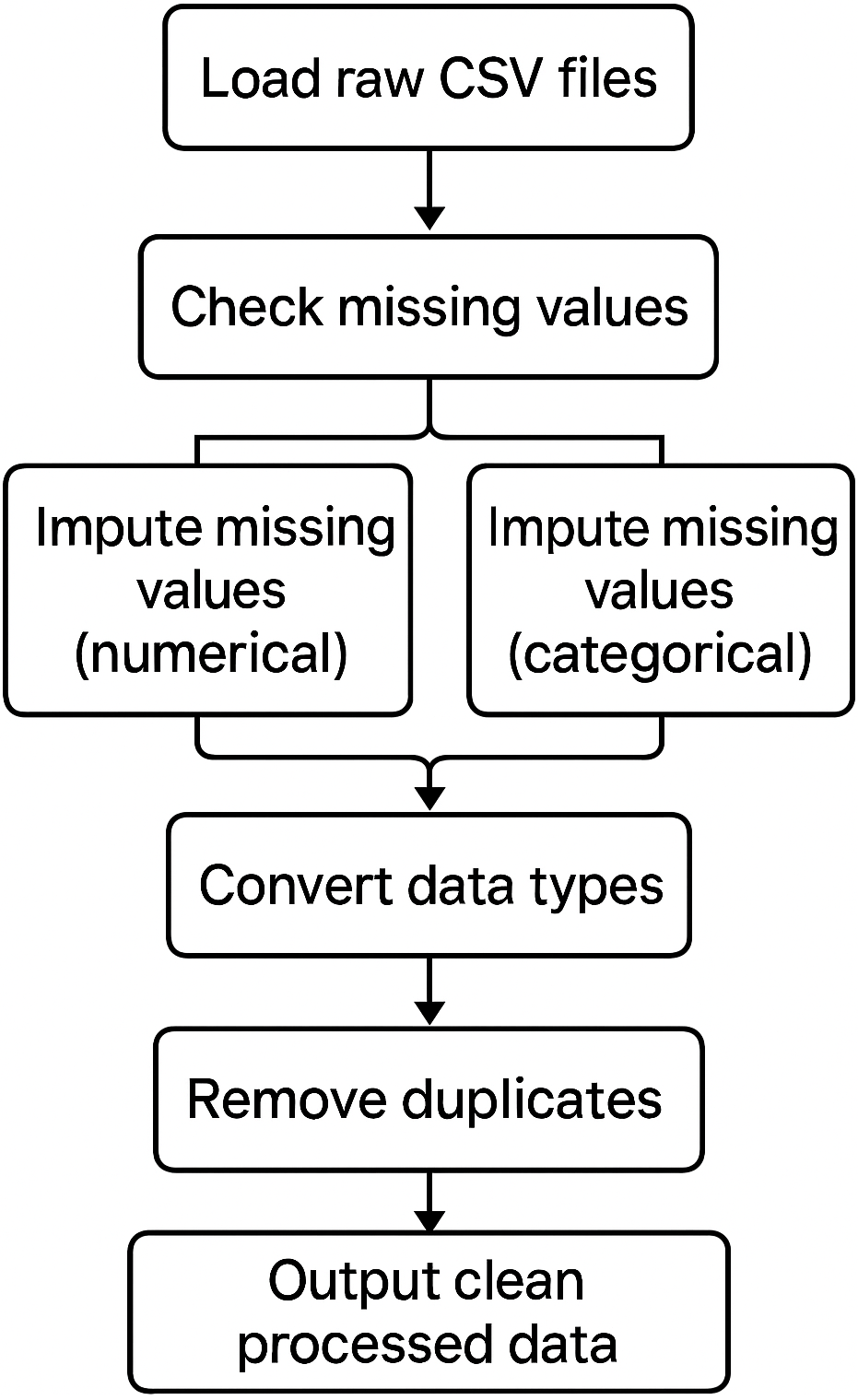

Figure 1: Data preprocessing pipeline.

2.3 Feature Engineering

Feature List

Feature	Description
Waste_Per_Capita_kg	(Waste/day × 1000)/population density.
Landfill_Utilization_Ratio	Waste generated/landfill capacity.
Year_Sin / Year_Cos	Cyclic encoding for temporal seasonality.
Waste_Generated_log	Log transform for skewed distribution.
Cost_of_Waste_log	Log transform for skewed distribution.
Eff_Campaign_Interaction	Municipal efficiency × awareness campaigns.
City_WasteType_TE	Target encoding of city/waste-type combination.
Lat/Long	Geospatial mapping of landfill sites.
Cluster_Dummies	Region cluster one-hot encoding.

2.4 Model Selection

Models tested:

- Linear Regression, Ridge, Lasso,
- Random Forest,
- Gradient Boosters: XGBoost, LightGBM, CatBoost.

CatBoost was chosen for the final model due to the best RMSE and native handling of categorical variables.

2.5 Hyperparameter Tuning

Hyperparameter optimization with Optuna:

- `depth` $\in [4, 10]$, `learning_rate` $\in [0.01, 0.3]$,
- `l2_leaf_reg` $\in [1, 10]$, iterations up to 2000,
- 50 trials with 5-fold cross-validation and early stopping.

2.6 Deployment and Deliverables

- Final model saved in `/models`.
- Streamlit app with pages: *Predictor*, *Data Explorer*, *Analytics*, *Map*.
- Deliverables:
  - `requirements.txt` – dependencies,
  - `README.md` – setup & usage,
  - `predictions.csv` – submission predictions.
- **Deployment Steps (Render example):**
  1. Push code to a public GitHub repository with `requirements.txt`.
  2. Add a `Procfile` (e.g., `web: streamlit run src/app.py --server.port $PORT --server.address 0.0.0.0`).
  3. Create a new Web Service on Render linked to the repo (Python $\geq$ 3.9).
  4. Set environment variables if needed and deploy.
- **Deployed App:** https://waste-management-11.onrender.com/

3 Exploratory Data Analysis (EDA) Visualizations

3.1 Target Distribution

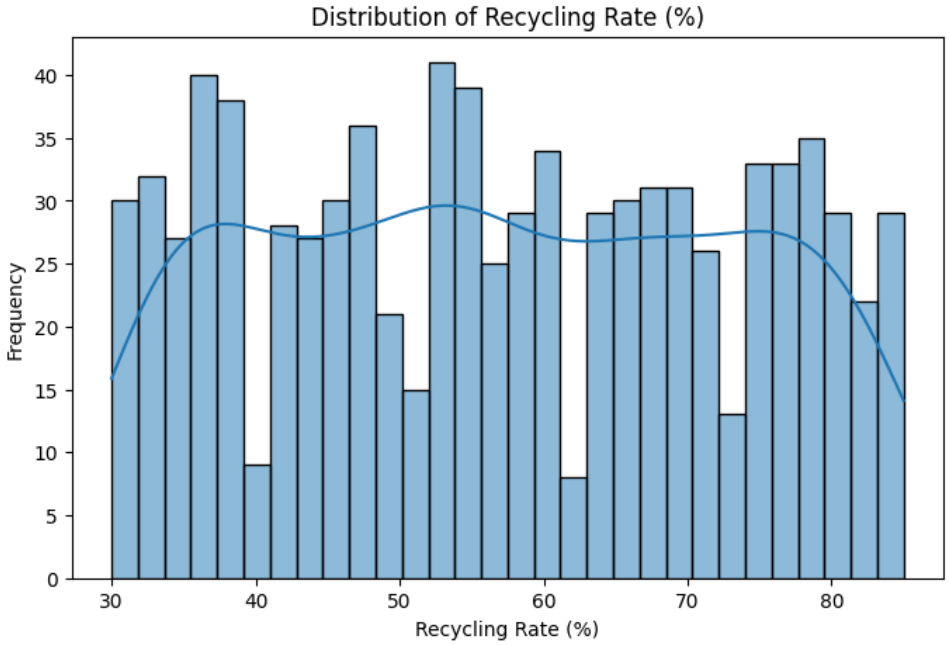

Figure 2: Distribution of Recycling Rate (%).

3.2 Feature Distributions

Histograms of major numeric features revealed skewness in waste generated and cost metrics—prompting log transforms.

3.3 Correlation Heatmap

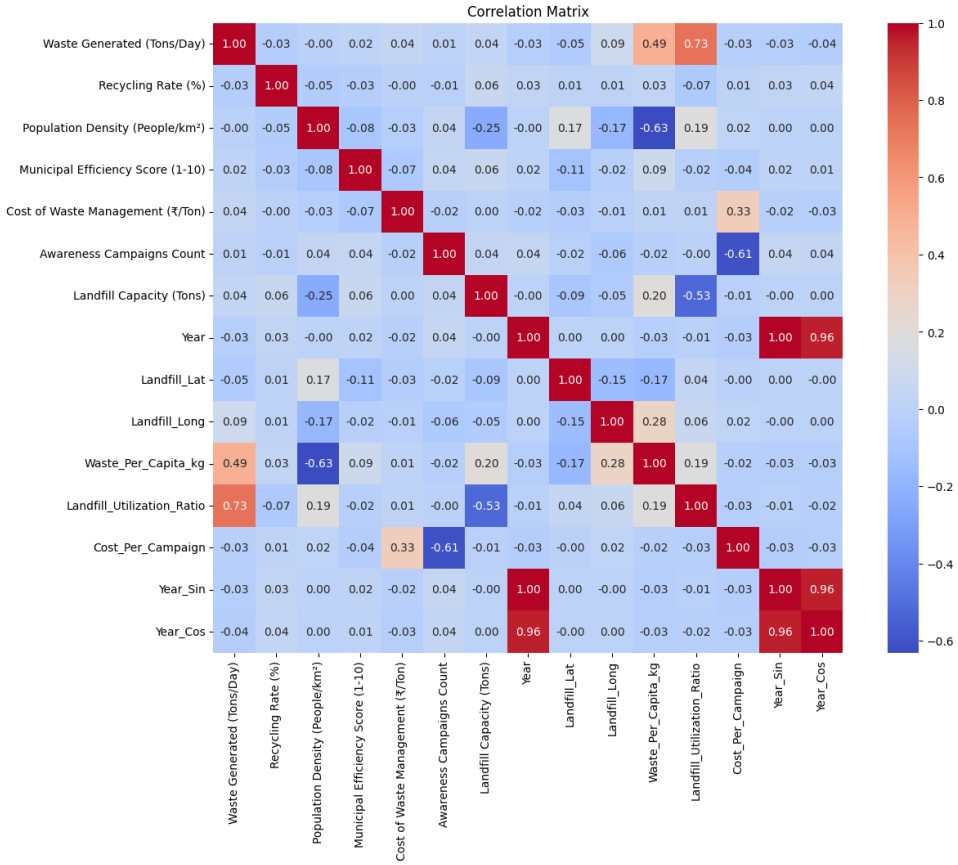

Figure 3: Correlation heatmap of numeric variables.

3.4 Categorical Comparisons

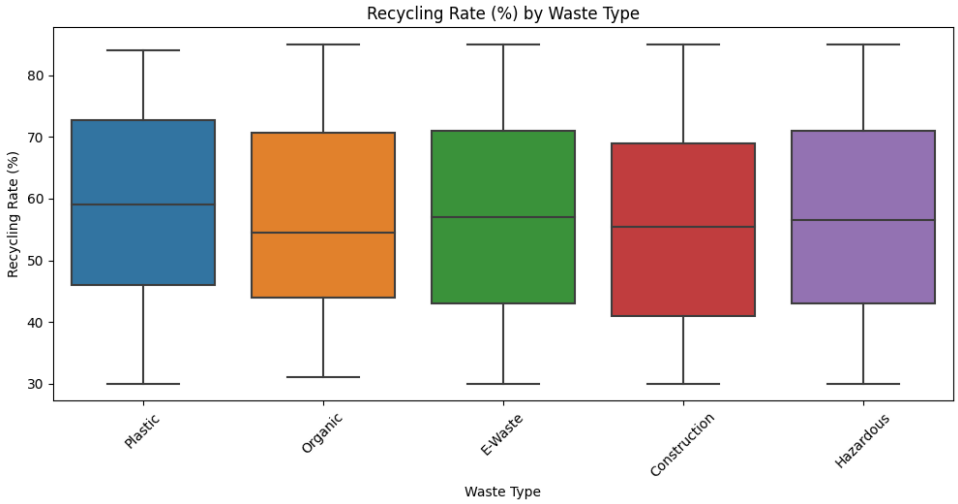

Figure 4: Recycling rate by waste type.

3.5 Geospatial View

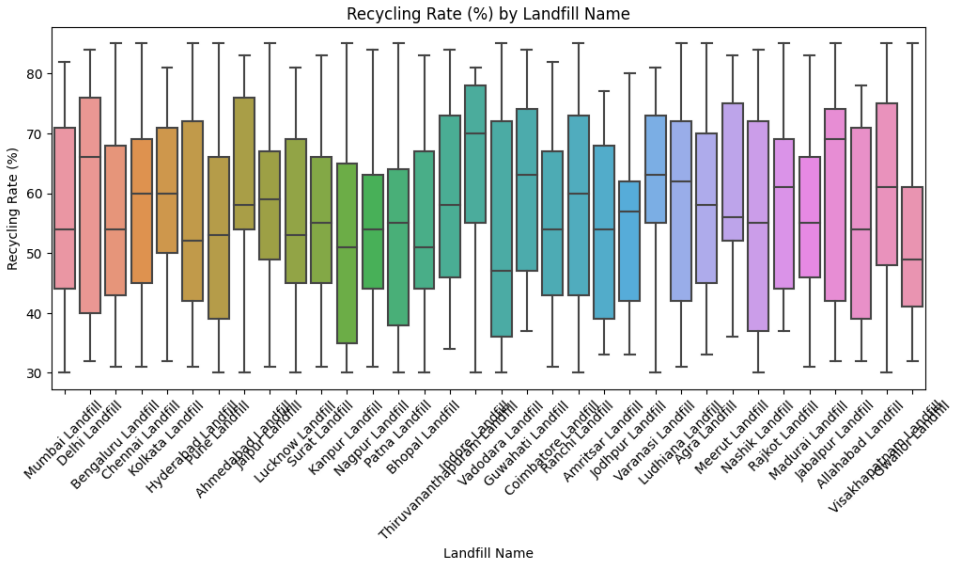

Figure 5: Landfill locations with utilisation ratios.

3.6 Average Recycling Rate

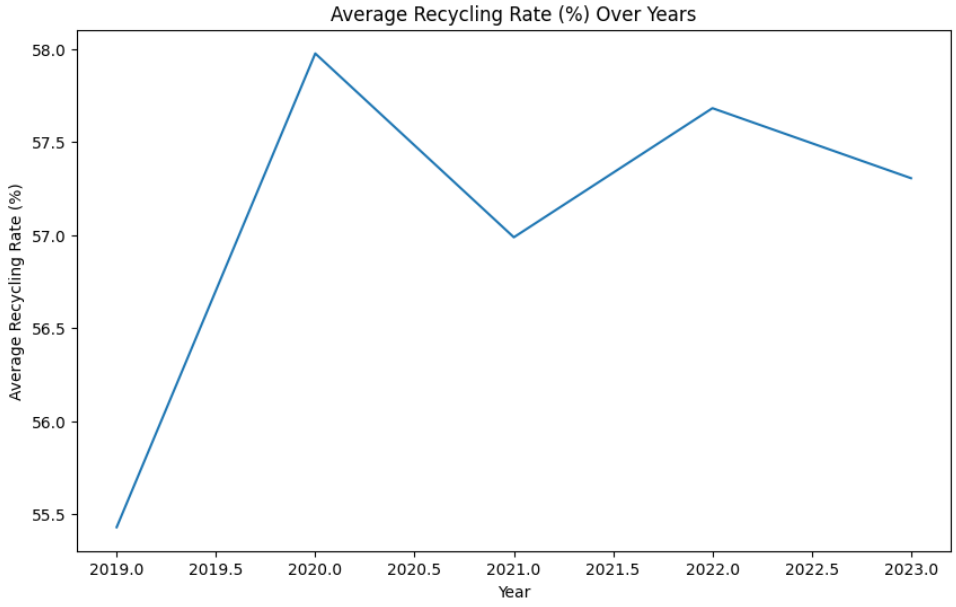

Figure 6: Average Recycling Rate (%) over years.

4 Results

Key Insights

- **CatBoost achieved the best RMSE of 13.96**, outperforming all baselines.
- Engineered features (e.g., *Waste_Per_Capita_kg*, *Eff_Campaign_Interaction*) contributed strongly to predictive power.
- Cities with higher municipal efficiency and active awareness campaigns tend to show higher recycling rates.

4.1 Model Performance

Table 2: Model Performance Comparison

Model	RMSE	R^2
Linear Regression	21.45	0.05
Random Forest	16.82	0.18
XGBoost	15.37	0.21
LightGBM	15.29	0.22
CatBoost	**13.9591**	**0.2336**

Takeaway: CatBoost delivered the lowest RMSE (13.96), demonstrating the advantage of gradient boosting with native categorical handling for this problem.

4.2 Feature Importance

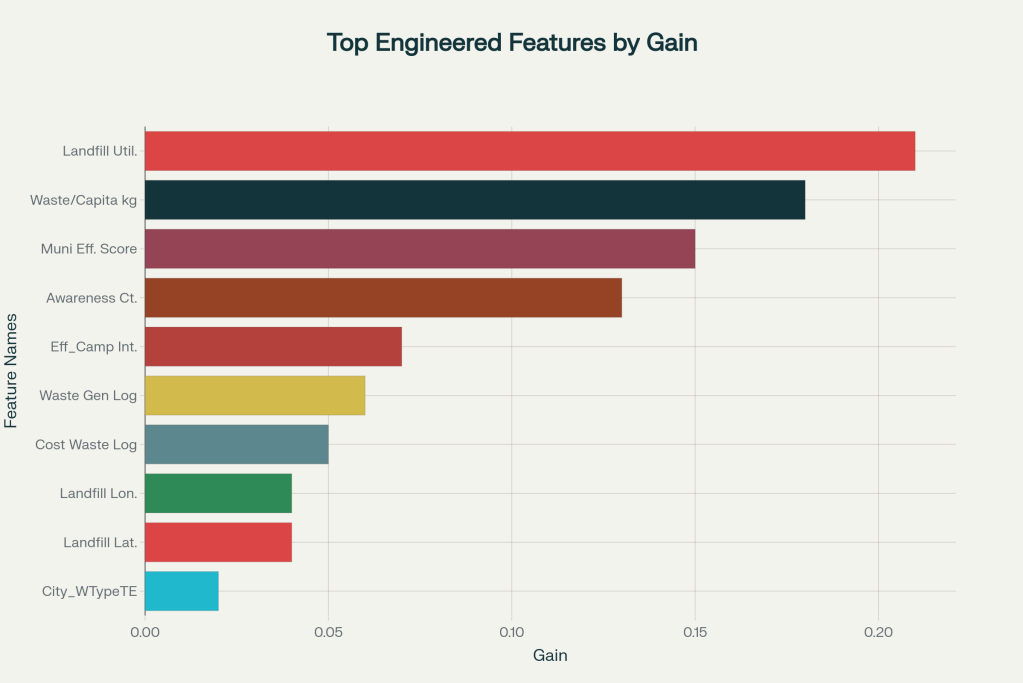

Figure 7: Top engineered features by CatBoost gain metric.

5 Discussion

5.1 Challenges

- Feature mismatch between training and prediction datasets.
- Cloud deployment path issues.
- Missing campaign count data in some records.

5.2 Limitations

- Simulated dataset; lacks ward-level granularity.
- Short time span—limited temporal modeling capacity.

5.3 Real-world Impact

- Informing municipal recycling policies and budgeting.

- Optimising landfill capacity and fleet routing.
- Guiding awareness campaign deployment.
- **Example:** Cities with higher municipal efficiency scores and more than 10 annual awareness campaigns achieved up to **15%** higher recycling rates than similar cities with fewer campaigns—suggesting a policy lever for targeted engagement.

5.4 Future Scope

Integration with IoT sensors, expansion to other regions, and deep learning approaches for time-series predictions. Additionally, **real-time integration with municipal data APIs** can enable live monitoring of recycling rates, and a **policy simulation dashboard** can help planners test the impact of interventions before implementation.

6 Conclusion

The project successfully predicted recycling rates with an RMSE of 13.96 using a tuned CatBoost model. Integration into a user-friendly web app enhances utility for municipal decision-making, aligning waste management operations with sustainability goals.

Acknowledgement

Thanks to the PWSkills Hackathon organisers, mentors, peers, and the open-source community for tools such as Python, Pandas, Streamlit, and CatBoost.

Appendix

Environment Details

Python	3.9.13
CatBoost	1.2.2
Optuna	3.1.1
Streamlit	1.27.2
Pandas	1.5.3

Folder Structure

```
project_root/
 Notebooks/
    data_preparation.ipynb
    exploratory_data_analysis.ipynb
    feature_engineering.ipynb
    model_selection.ipynb
    model_training.ipynb
 data/
    raw/
    processed/
 models/
    catboost_tuned_model.cbm
    catboost_tuned_model.pkl
 src/
    data/
    models/
    utils/
 requirements.txt
 static/
 templates/
 README.md
 predictions.csv
 report.pdf
```

Sample predictions.csv

City	Year	Predicted Recycling Rate
Agra	2024	66.91304
Chennai	2027	61.04348
Agra	2029	61.00000
Agra	2030	64.00000
Agra	2030	43.56000
Visakhapa	2030	56.66667
Thiruvanar	2027	62.66667
Thiruvanar	2027	61.73913
Bhopal	2026	64.21739
Bengaluru	2026	52.00000
Bengaluru	2026	58.28000
Bengaluru	2027	51.00000
Bengaluru	2027	50.30435
Bengaluru	2029	54.95652
Bengaluru	2029	78.00000
Jabalpur	2029	58.28000
Nagpur	2026	41.95652
Kolkata	2026	56.86957

References

- CatBoost Documentation – https://catboost.ai/
- Streamlit Documentation – https://streamlit.io/
- Optuna – https://optuna.org/