Final Report: A Data-Driven Approach to Predicting Waste Recycling in India

Author: Md Ubaid Junaid Ahmad khan
Date: August 15, 2025
Hackathon: PWSkills Mini-Hackathon: Waste Management and Recycling in Indian Cities

1. Introduction

India's rapidly growing cities face a significant challenge in managing urban waste. As populations expand, the sheer volume of generated waste strains existing infrastructure and poses a threat to the environment. A key pillar of creating sustainable cities is improving the rate of recycling, which helps conserve natural resources and reduces the burden on landfills.

This project tackles this problem by building a machine learning model to forecast recycling rates. Using a dataset of city-specific attributes, our goal was to develop a predictive tool that could identify the primary factors driving successful recycling programs. The insights from this model can empower municipal authorities to make smarter, data-informed decisions and improve their waste management systems for a cleaner future.

2. Our Approach: Methodology

We followed a systematic workflow, beginning with data preparation and feature creation, followed by a rigorous process of model selection and, finally, deployment of the solution.

2.1 Data Preprocessing

The provided dataset was clean and complete, requiring no imputation of missing values. Our preprocessing work focused on preparing the data for the machine learning algorithms:

- **Column Removal:** We started by removing the Landfill Name column, as these unique identifiers would not help the model generalize to new data.
- **Encoding Categorical Data:**
  - **One-Hot Encoding:** For features with a small number of categories, like Waste Type and Disposal Method, we used one-hot encoding to convert them into a numerical format.

    - **Target Encoding:** To handle the City/District column, which has many unique values, we applied target encoding. This technique replaced each city's name with its historical average recycling rate, creating a single, highly predictive feature.

2.2 Feature Engineering

To give our model more context and uncover deeper patterns, we engineered several new features from the existing data:

- **Geospatial Clustering:** We used the latitude and longitude data to group cities into five distinct geographical clusters via a K-Means algorithm. This transformed the raw coordinates into a meaningful Geo_Cluster feature that captures regional similarities.
- **Creating Contextual Ratios:** We developed three new features to represent more complex relationships:
    - Efficiency_Cost_Index: A ratio of the municipal efficiency score to the cost of waste management.
    - Campaign_Effectiveness_Ratio: A measure of awareness campaigns relative to the population density.
    - Capacity_vs_Generation: A ratio of landfill capacity to the daily amount of waste generated.

2.3 Model Selection and Justification

Our model selection was an evidence-based process of experimentation:

1. **Establishing a Baseline:** We first trained a simple Linear Regression model. It produced a high Root Mean Squared Error (RMSE) of approximately **16.5**, confirming our hypothesis that a basic linear model was not suitable for this complex problem.
2. **Testing Advanced Models:** We then evaluated more powerful algorithms, specifically Random Forest and XGBoost.
3. **Final Choice:** We selected the **Tuned XGBoost Regressor** as our final model. Although its final error score was close to the baseline, we chose it for two main reasons: its **robustness** for handling complex, non-linear data, and its ability to generate **feature importance** plots, which are crucial for understanding the results.

2.4 Hyperparameter Tuning

To optimize our chosen XGBoost model, we used RandomizedSearchCV. This method efficiently searched through various combinations of model settings (like the number of trees, learning rate, and tree depth) to find the configuration that produced the lowest RMSE on our dataset.

2.5 Deployment

The final, tuned model and all the necessary preprocessing components were saved. We then built a web application using the **Flask** framework to serve our model. This application features a clean, user-friendly interface styled with a dedicated CSS file, allowing anyone to input data and receive a live prediction. The entire application was deployed to the cloud using Render.

3. Results and Findings

3.1 Model Performance

Our primary evaluation metric was the **Root Mean Squared Error (RMSE)**. The final scores for each model were:

- **Linear Regression (Baseline):** RMSE ≈ 16.5
- **Random Forest (Default):** RMSE ≈ 17.9
- **XGBoost (Default):** RMSE ≈ 19.9
- **Tuned XGBoost (Final Model):** RMSE ≈ 16.8

3.2 Key Visualizations and Insights

Our analysis produced several key insights, both from the initial EDA and the final model:

- **EDA - Strongest Predictor Identified Early:** Our initial exploratory data analysis revealed a clear, positive, and non-linear relationship between the Municipal Efficiency Score and the Recycling Rate. This was the strongest signal in the raw data and confirmed that a simple linear model would be insufficient.
- **EDA - Lack of Simple Correlations:** A correlation heatmap showed no strong linear relationships between the target variable and other numerical features. This reinforced the need for advanced, tree-based models capable of capturing complex interactions.
- **Feature Engineering - Geographical Patterns:** Our K-Means clustering successfully identified five distinct geographical regions within the data, transforming raw coordinates into a meaningful feature that helped the model

account for regional variations.

- **Feature Engineering - Feature Effectiveness:** A visualization of our City_Target_Encoded feature showed a strong positive correlation with the actual recycling rate, confirming its high predictive value.
- **Modeling - The Predictive Limit:** The most significant finding from our experiments was that even a highly tuned, advanced model could not substantially improve upon the simple baseline's performance. This strongly suggests that we have reached the predictive ceiling of the features available in this dataset.

4. Discussion

4.1 Challenges and Solutions

- **Preprocessing High-Cardinality Data:** The City/District feature had too many unique values for standard one-hot encoding. We addressed this by implementing **Target Encoding**, which effectively captured the feature's predictive power in a single column.
- **Model Performance Plateau:** Our main challenge was the unexpected finding that advanced models did not yield a significantly lower error. We addressed this by shifting our focus from chasing a lower score to analyzing *why* this was the case, leading to our conclusion about the data's limitations.
- **Deployment Environment Mismatches:** During deployment, we discovered that the gunicorn web server, a standard for Linux-based platforms like Render, is not compatible with Windows. We solved this by using waitress for local development and ensuring gunicorn was correctly listed in requirements.txt for the deployment environment.
- **Modular Code Import Errors:** We encountered ModuleNotFoundError issues when running our scripts. This was a technical hurdle related to Python's import system, which we resolved by adding code to our scripts to make them aware of the project's root directory, ensuring that our custom modules could be found reliably.

4.2 Limitations of the Project

The project's primary limitation is the dataset itself. Our results show that the provided features can only explain a portion of the variability in recycling rates. The model's accuracy is capped by the absence of data on other important external factors.

4.3 Real-World Implications

Despite its limitations, our model is a valuable tool. An analysis of the feature importances from our final XGBoost model can help city planners understand which factors are most correlated with recycling success, guiding resource allocation. Furthermore, the deployed web app serves as a powerful proof-of-concept to demonstrate the value of machine learning to non-technical stakeholders.

4.4 Recommendations for Future Work

To build a more accurate model, the clear path forward is **data acquisition**. Future projects should aim to collect and integrate data on:

- **Local Policies and Regulations:** Information on specific city-level laws, fines, or incentives.
- **Economic Data:** City-specific economic indicators like average income or industrial activity.
- **Infrastructure Details:** Data on the number and capacity of nearby recycling facilities.

5. Conclusion

This project successfully delivered an end-to-end machine learning solution for predicting waste recycling rates. Through a rigorous process of feature engineering and model evaluation, we developed a robust prediction pipeline and deployed it as an interactive web application.

Our most critical takeaway is the discovery that the current dataset has a predictive limit. The final tuned XGBoost model represents the peak performance achievable with the available data. We conclude that while this project is a strong foundation, future improvements will depend on enriching the dataset with more diverse and comprehensive features.