SOMAIYA
VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

EnviroMeter - Carbon Footprint Predictor

Dishita Shah – 16010122167 || Aakanksh Sen - 16010122166
Krish Satra – 16010122164 || Heemit Shah - 16010122168

Guided By:- Dr. Shila Jawale

Motivation

- Climate change is one of the most urgent global challenges, yet most individuals are unaware of how their daily lifestyle choices contribute to carbon emissions.
- People struggle to estimate their personal environmental impact because carbon footprint calculations are complex and depend on many factors like transport, diet, energy usage, and consumption habits.
- Existing tools are either too generalized or require domain expertise, making them inaccessible for regular users or students.
- There is a need for a simple, intelligent, data-driven system that can accurately estimate a person's carbon footprint and give personalized recommendations for reducing it.
- With the rise of machine learning, we can analyze real lifestyle data and build a smart predictor that helps individuals make environmentally responsible decisions.
- This project aims to empower users with awareness, make sustainability measurable, and encourage small changes that lead to a significant positive impact on the planet.

Problem Statement

The Problem:

- Individuals lack awareness of how their daily activities such as travel, food consumption, and energy usage contribute to their overall carbon footprint.
- Existing carbon footprint calculators are often complex, time-consuming, and provide generalized estimates that are not tailored to personal lifestyle patterns.
- There is no simple, accessible tool that offers personalized and actionable insights to help users reduce their environmental impact.

Our Solution:

EnviroMeter: An AI-powered carbon footprint predictor that analyzes user behavior and delivers personalized carbon footprint assessments along with practical, data-driven recommendations for sustainable living.

Objectives

The objectives of our project are as follows:

1. To develop an AI-driven model that accurately predicts an individual's carbon footprint based on lifestyle inputs.
2. To provide users with personalized, easy-to-understand insights about their environmental impact.
3. To recommend actionable steps and sustainable habits that help users reduce their carbon footprint.
4. To create a user-friendly interface that allows seamless data entry and real-time footprint calculation.
5. To promote environmental awareness by visualizing carbon emissions across different lifestyle categories (travel, energy, food etc.).

Comparative Overview – Summary, Strengths & Limitations

Aspect	Paper 1	Paper 2	Paper 3	Paper 4
Summary	Carbon-aware optimization of USDA Thrifty Food Plan using DBSCAN clustering (~11.3% reduction).	Triple-ensemble (RF, CatBoost, DNN) to predict CO2 from behavioral & vehicle data with SHAP explanations.	Predicts emissions from campus activities (commuting, energy use) and deploys web dashboard.	Scalable web app for personalized footprint estimation and recommendations.
Strengths	Policy-ready; cost & nutrition retained; interpretable clustering.	High accuracy; explainable; robust ensemble.	Uses real data; actionable; promotes awareness.	User-focused; modular; engaging design.
Limitations	Cradle-to-gate only; assumes user adoption; regional recalibration needed.	Synthetic data limits realism; overfitting risk; high compute.	Site-specific; survey bias; data-frequency issues.	Depends on user input accuracy; needs regular emission updates.

Comparative Overview – Methodology, Datasets & Evaluation Metrics

Aspect	Paper 1	Paper 2	Paper 3	Paper 4
Methodology	Optimization + DBSCAN clustering on food categories to minimize carbon.	Stacked ensemble of RF, CatBoost, DNN with SHAP-based explainability.	Regression + LSTM; fuses commuting, energy, and material usage data.	Backend ensemble model with rule-based and SHAP explanation modules.
Datasets	WWEIA, NHANES, FNDDS, FPED, DataFRIENDS carbon data.	Vehicle CO2 datasets + synthetic human footprint data (~10k).	ITB University campus data (electricity, surveys, travel).	User-input lifestyle data + global emission factor libraries.
Evaluation Metrics	% Carbon reduction, cost, nutrition constraints.	R^2, MAE, RMSE, CV-R^2, SHAP importance.	R^2, RMSE, MAE, usability satisfaction (SUS=4.2).	R^2, RMSE, MAE, engagement rate, completion rate.

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

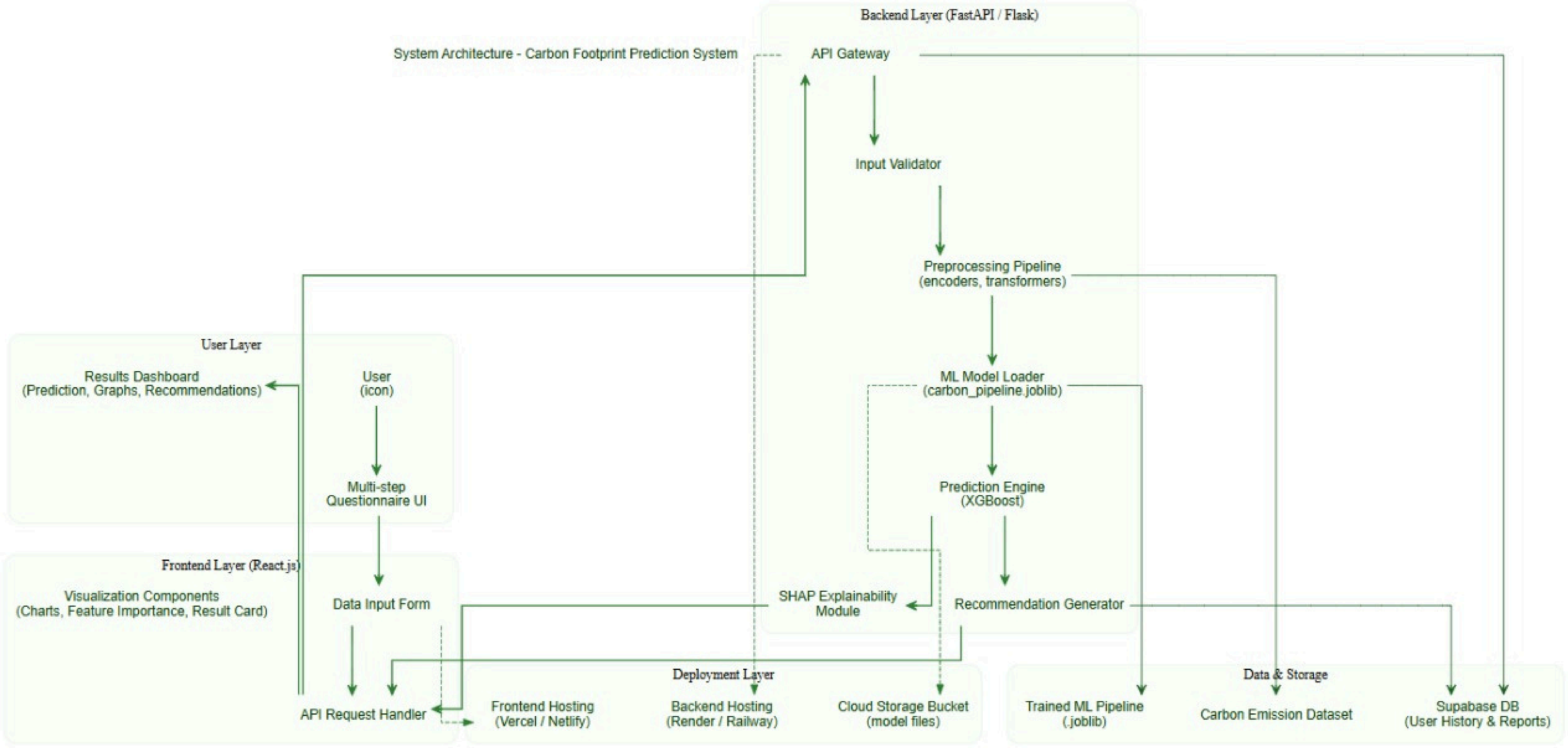
Comparative Overview – Results, Discussion & Impact

Aspect	Paper 1	Paper 2	Paper 3	Paper 4
Results	11.3% carbon reduction; maintained cost & nutrition.	R^2=99.7% (vehicle), 98.4% (human) – high accuracy.	612.8 tCO_2e/year; per-student 571 kg/year.	High user usability (4.5/5) & engagement.
Discussion	Low-carbon diet viable in national plans.	SHAP identified behavioral factors; interpretable.	Temporal models useful for sustainability monitoring.	Personalized insights increase eco-awareness.
Impact	Can influence policy-level nutrition frameworks.	Reliable for policymaking & behavioral studies.	Helps institutions plan emission reduction strategies.	Motivates individual behavioral change.

Comparative Overview – Applications & Improvements

Aspect	Paper 1	Paper 2	Paper 3	Paper 4
Applications	Dietary policy redesign for low-carbon menus.	Fleet & lifestyle CO2 monitoring tools.	Campus dashboards for emission visualization.	Personalized carbon footprint awareness apps.
Scalability	Applicable nationwide; can integrate into USDA programs.	Extensible across regions and datasets.	Needs replication at multiple institutions.	Highly scalable due to modular web backend.
Future Improvements	Add cradle-to-grave analysis & behavioral modeling.	Use real-world datasets to reduce synthetic bias.	Expand to cross-campus comparisons.	Automate regional emission calibration & updates.

System Architecture



Dataset Description

Dataset Overview

- **Source:** Kaggle – *Individual Carbon Footprint Calculation Dataset*
- **Type:** Synthetic dataset generated from aggregated studies on lifestyle, transport, and energy habits.
- **Purpose:** To estimate total personal carbon emissions (in kg CO_2e) based on behavioral and lifestyle attributes.
- **Total Features:** 19 independent variables + 1 target variable (CarbonEmission).
- **Nature of Data:**
  - *Mixed-type* — includes categorical (e.g., diet, transport, recycling) and numerical (e.g., grocery bill, distance) features.
  - *Representative, not real* — simulates realistic household behavior patterns using statistically weighted distributions.

Key Features

- **Lifestyle Factors:** Diet, Body Type, Frequency of Shower, Social Activity, New Clothes Purchased, TV/Internet Usage.
- **Energy Use:** Heating Energy Source, Energy Efficiency Preference, Cooking Devices.
- **Transportation:** Transport Mode, Vehicle Type, Vehicle Distance per Month, Air Travel Frequency.
- **Waste Generation:** Waste Bag Size & Weekly Count, Recycling Habits.
- **Expenditure Indicators:** Monthly Grocery Bill.
- **Target Variable:** CarbonEmission → total personal CO_2 equivalent output per individual.

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Preprocessing Summary

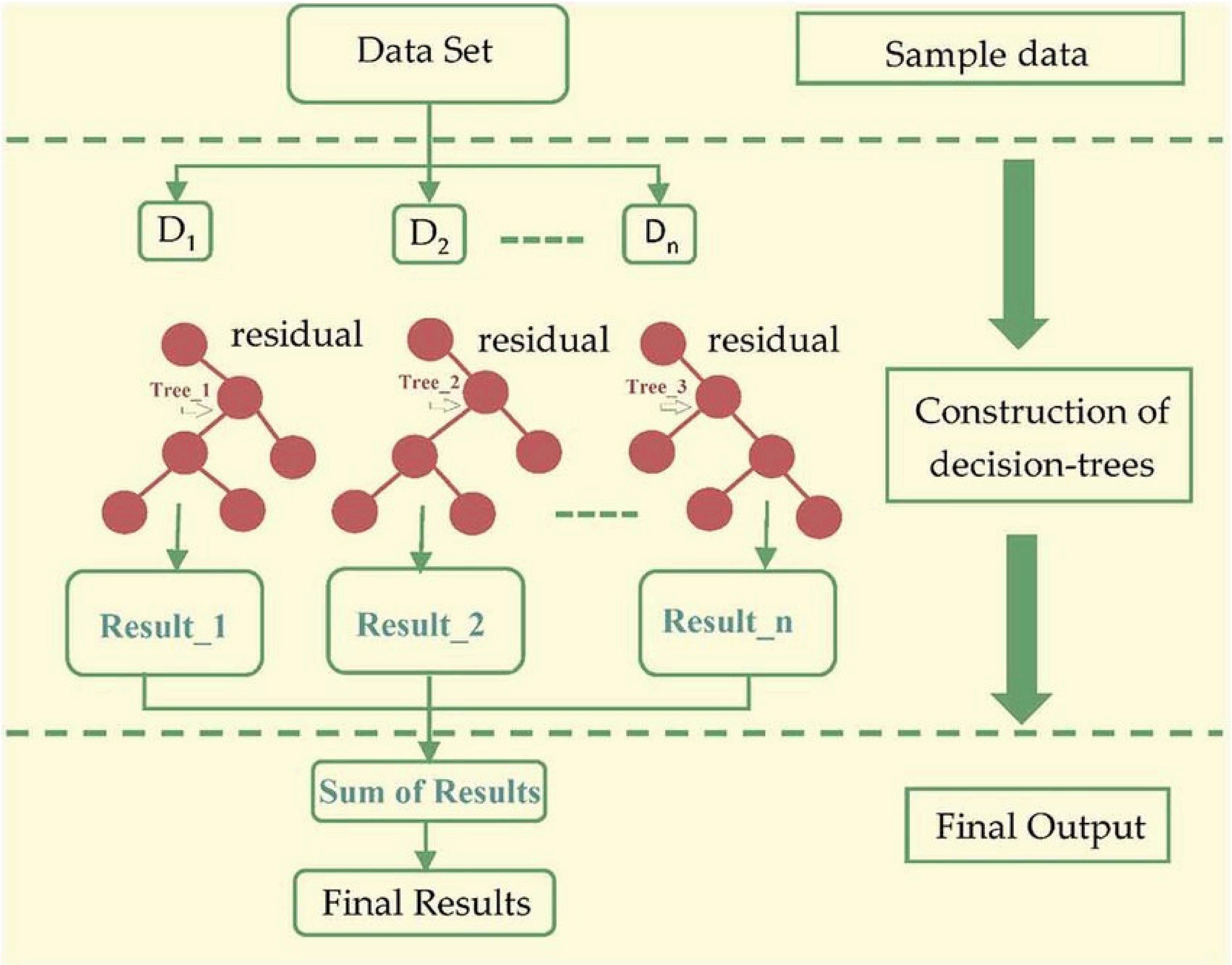
Before model training, data underwent systematic **feature engineering and transformation** steps to ensure compatibility and consistency:

- **Missing Value Handling**
  - Checked and imputed missing or inconsistent values.
  - Verified feature completeness since data was synthetically generated.
- **Feature Categorization**
  - Split all columns into **numerical** (e.g., distance, bills, usage hours) and **categorical** (e.g., transport, diet, recycling).
- **Scaling of Numerical Features**
  - Applied **standard normalization** to make features comparable and stabilize gradient-based algorithms.
- **Encoding of Categorical Features**
  - Converted textual categories (e.g., “Diesel”, “Electric”, “Vegetarian”) into **machine-readable vectors** using **One-Hot Encoding**.
- **Unified Transformation Pipeline**
  - Integrated both numeric and categorical transformations via a **ColumnTransformer**, ensuring a single, consistent preprocessing step for the full dataset.
- **Final Dataset Shape**
  - Approximately **10,000 samples × ~60 transformed features** after encoding and scaling.

XGBoost — Model Architecture & Hyperparameters

XGBoost architecture (tree boosting ensemble):

- Ensemble of gradient-boosted decision trees. Each tree fits residuals of the previous ensemble.
- Handles heterogeneous features, missingness, and interactions automatically.



XGBoost — Training & Evaluation Pipeline

Training Configuration

- **Boosting Rounds:** 300 trees (balanced between learning stability and training time)
- **Learning Rate:** 0.05 → slower, more stable convergence
- **Maximum Depth:** 6 → controls model complexity and avoids overfitting
- **Subsampling:** 0.8 → uses random samples per tree for better generalization
- **Column Sampling:** 0.8 → prevents dominance of specific features
- **Evaluation Metric:** Root Mean Squared Error (RMSE)
- **Objective Function:** Minimize squared error (reg:squarederror)
- **Early Stopping:** Stops if validation RMSE doesn't improve after 30 rounds

Training Process

- **Data Split:** Dataset divided into 80% training and 20% validation sets.
- **Model Fitting:** XGBoost trained iteratively, learning residual patterns between predicted and actual emissions.
- **Validation Monitoring:** Performance monitored on validation data to detect overfitting early.
- **Feature Importance Extraction:** After training, the most influential lifestyle features were analyzed.

Evaluation Strategy

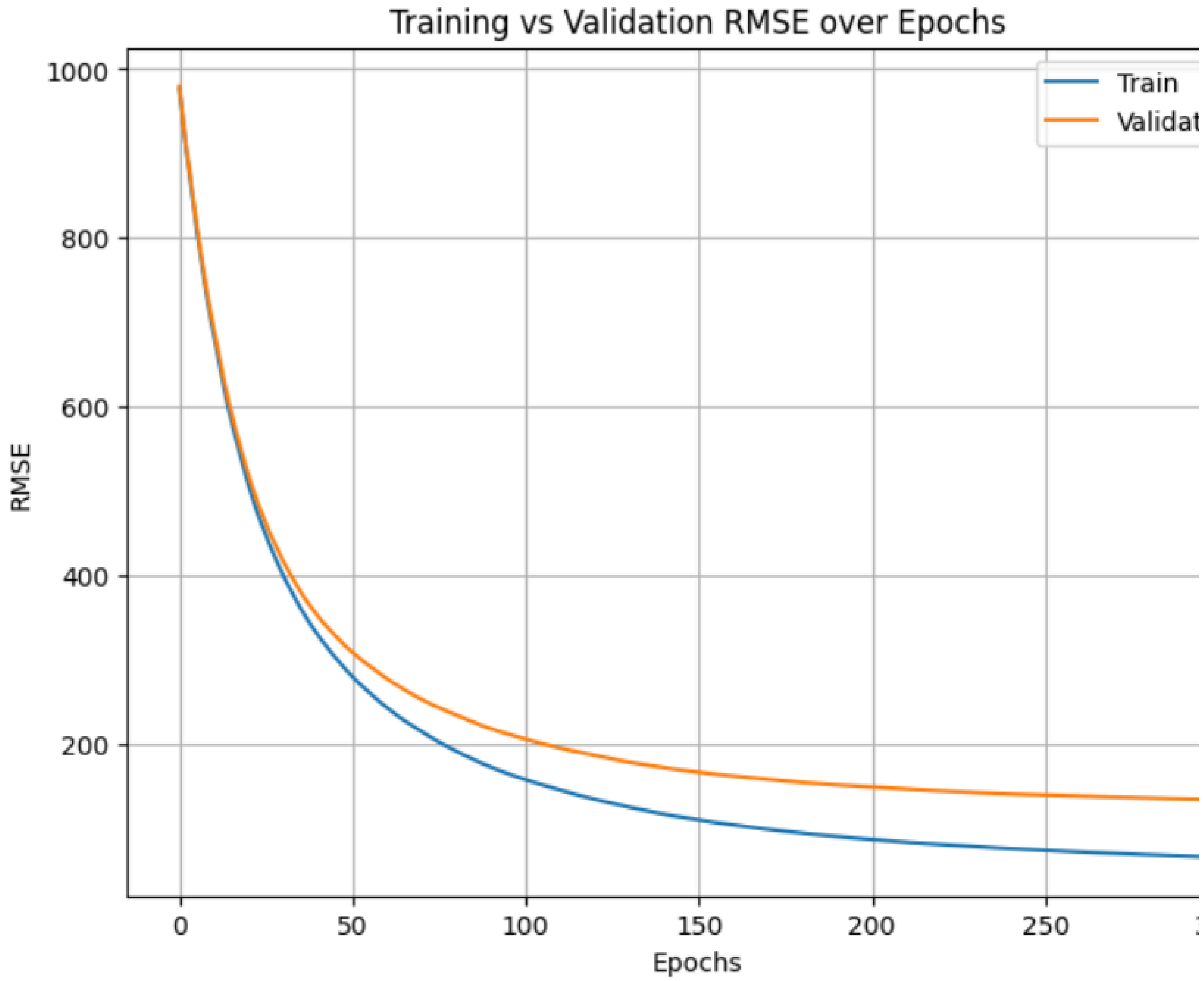
- **Primary Metrics:**
  - **MAE (Mean Absolute Error):** Measures average prediction error.
  - **RMSE (Root Mean Squared Error):** Penalizes large deviations more heavily.
  - **R^2 Score:** Represents how well the model explains variance in carbon emissions.
- **Cross-Validation:**
  - 5-fold validation confirmed model consistency across splits.
  - Average Cross-Validation $R^2 \approx$ **0.983 ± 0.001**

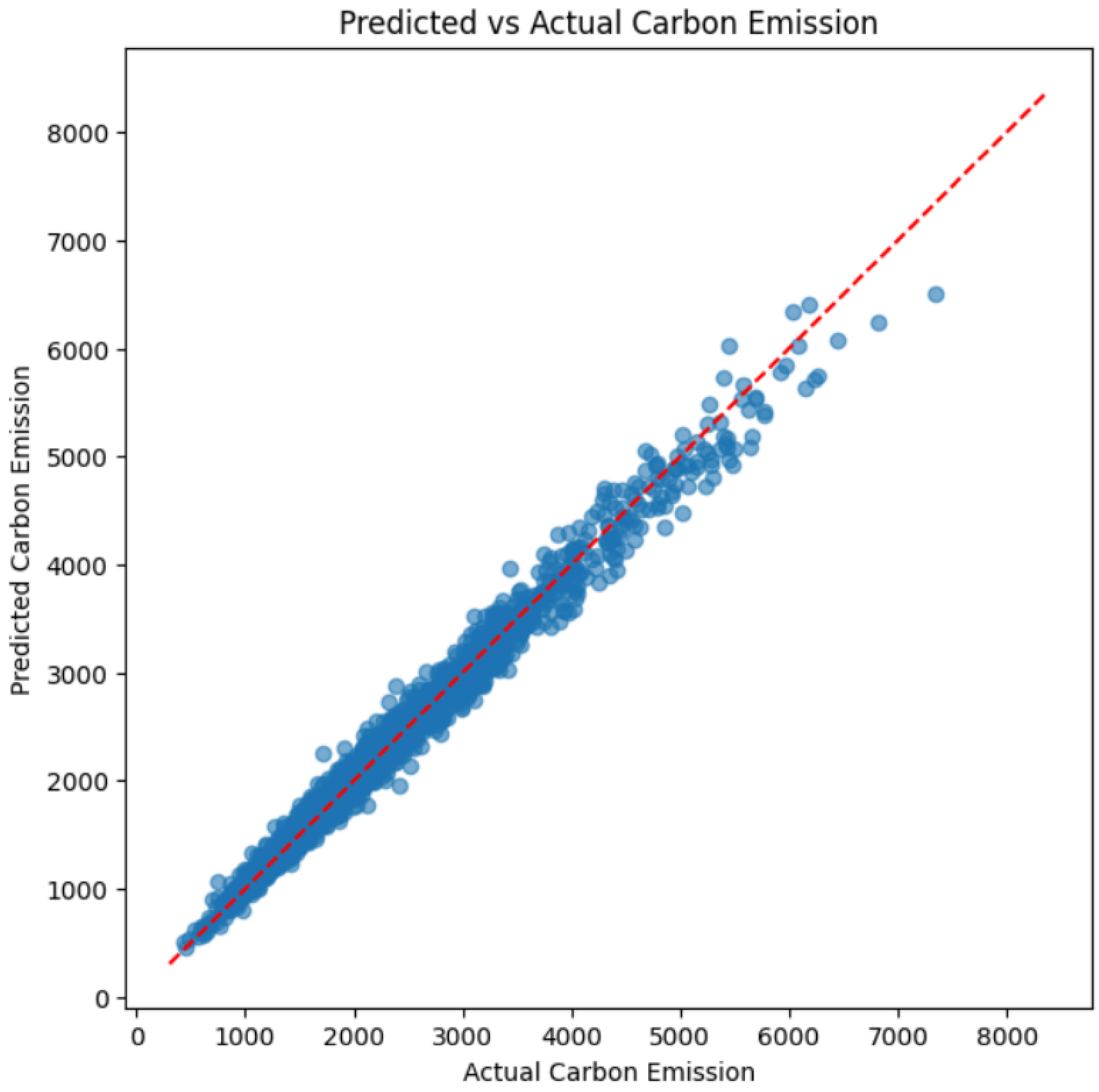
XGBoost — Results & Graphs

Model performance metrics (XGBoost):

- MAE: 94.621
- RMSE: 129.487
- R^2: 0.984

- Cross-Validation R^2: 0.983 ± 0.000
  Interpretation:
- The high R^2 indicates strong predictive capability; residual analysis should be inspected to confirm homoscedasticity.
- MAE and RMSE are in the units of the dataset (CO2e units) and show typical prediction errors - useful to map to real-world kgCO2e impacts.

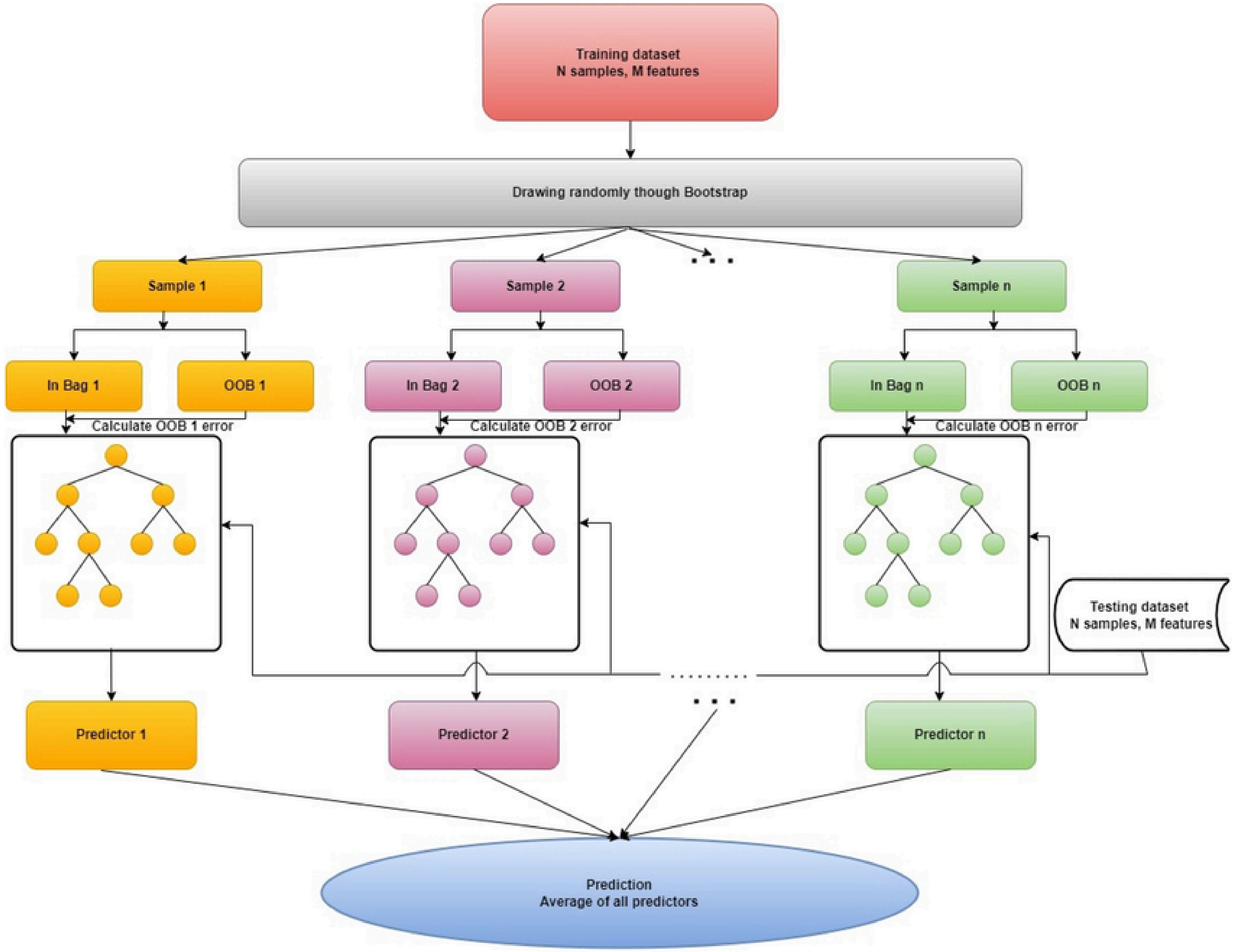

Training vs Validation RMSE over Epochs
1000
800
600
400
200
RMSE
0
50
100
150
200
250
Epochs
Train
Validat


Predicted vs Actual Carbon Emission
8000
7000
6000
5000
4000
3000
2000
1000
0
Predicted Carbon Emission
0
1000
2000
3000
4000
5000
6000
7000
8000
Actual Carbon Emission

CatBoost — Model Architecture & Hyperparameters

CatBoost architecture (tree boosting ensemble):

- Gradient boosting on decision trees with ordered boosting to reduce target leakage in categorical encodings.
- Native categorical feature handling reduces need for one-hot encoding and preserves ordinal information where present.



CatBoost — Training & Evaluation Pipeline

Training Configuration

- **Iterations:** 500 boosting rounds (trees) to ensure thorough convergence.
- **Learning Rate:** 0.05 → balances learning speed and stability.
- **Tree Depth:** 8 → allows moderate model complexity to capture nonlinear interactions.
- **Loss Function:** Root Mean Squared Error (RMSE) → focuses on minimizing squared deviations.
- **Evaluation Metric:** RMSE on validation data to monitor overfitting.
- **Random Seed:** 42 for reproducibility.
- **Early Stopping:** Training stops automatically if validation RMSE does not improve for 30 consecutive rounds.
- **Verbose Logging:** Progress printed every 50 iterations for transparency.

Training Process

- **Data Split:** 80% training, 20% validation to ensure unbiased evaluation.
- **Native Categorical Handling:**
  - Categorical feature indices were passed directly to the model (cat_features), allowing **CatBoost** to internally encode categories efficiently using **ordered target statistics**.
- **Boosting Mechanism:**
  - Each new tree is built to correct the residuals from previous trees, with leaf values optimized using ordered boosting.
- **Regularization:**
  - Controlled through learning rate and early stopping to maintain generalization.

Evaluation Strategy

- **Performance Metrics:**
  - **MAE (Mean Absolute Error):** Measures the average deviation between predicted and actual values.
  - **RMSE (Root Mean Squared Error):** Highlights larger prediction errors more strongly.
  - **R^2 Score:** Indicates the proportion of variance in emissions explained by the model.
- **Validation:**
  - Model performance evaluated on a held-out validation set.
  - Stable convergence observed around 450 iterations, confirming robust fit without overfitting.

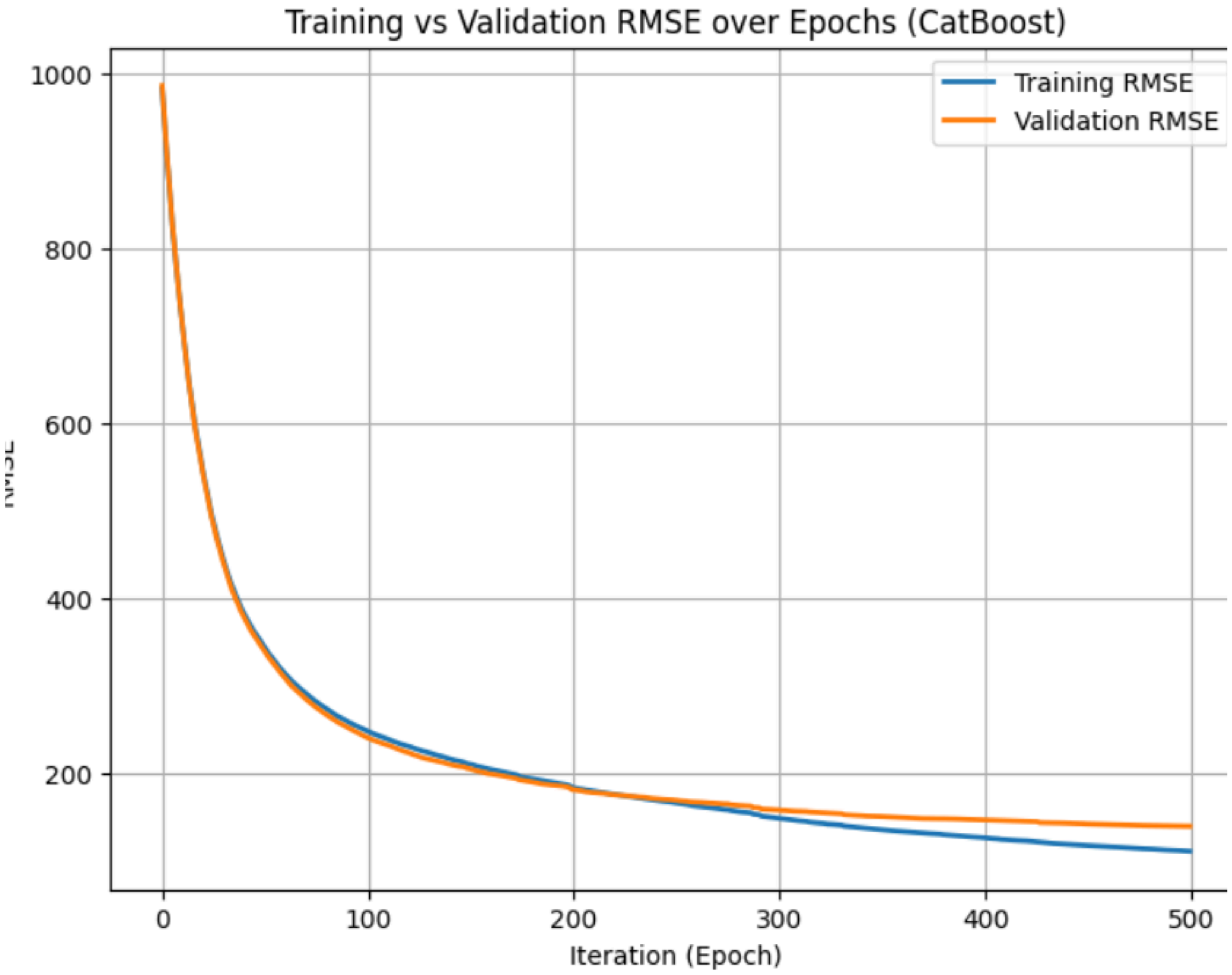
CatBoost — Results & Graphs

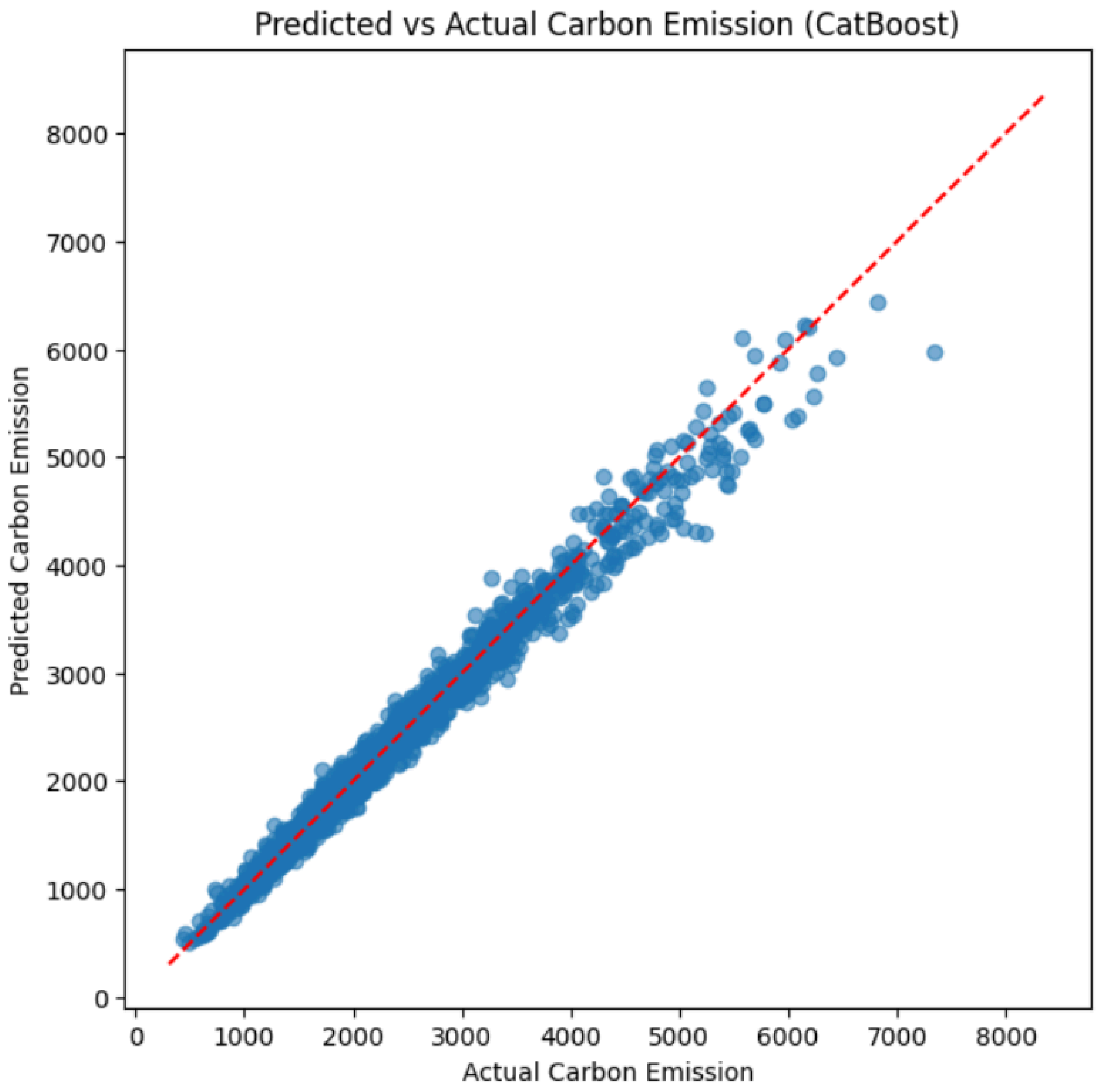
Model performance metrics (XGBoost):

- MAE: 96.445
- RMSE: 139.964
- R^2: 0.981

Interpretation:

- CatBoost produced competitive results with slightly higher RMSE but still high R^2 indicating reliable prediction.


Training vs Validation RMSE over Epochs (CatBoost)
Training RMSE
Validation RMSE
1000
800
600
400
200
0
100
200
300
400
500
Iteration (Epoch)


Predicted vs Actual Carbon Emission (CatBoost)
Predicted Carbon Emission
Actual Carbon Emission
0
1000
2000
3000
4000
5000
6000
7000
8000

Conclusion

Key takeaways:

- Combining optimization, explainable ML, and user-centered design enables practical carbon reduction solutions across policy, institutions and individuals.
- High predictive accuracy (ensemble models) must be paired with real-world data and uncertainty quantification to be trusted in policy settings.
- Future work should focus on generalization, causal testing of interventions, and closed-loop evaluation where recommendations lead to measurable emission reductions.

THANK YOU