**Carbon Footprint Prediction Project Report**

**1. Project Overview**

This project aims to predict an individual's carbon footprint using machine learning techniques, with a focus on LightGBM and PyCaret. It leverages various lifestyle and consumption features to estimate household carbon emissions with high accuracy.

**2. Dataset Description**

The dataset includes the following features:

* Electricity consumption (kWh/month)
* Natural gas consumption (therms/month)
* Vehicle miles traveled
* House area (sqft)
* Water usage (liters/day)
* Diet type and meat consumption (kg/week)
* Recycling and composting habits
* Insulation quality and use of smart appliances
* Laundry and transport behavior
* Heating type and household size

These features represent core factors that affect a household's carbon emissions.

**3. Data Preprocessing**

1. **Missing Values**:
   * Mode imputation for categorical variables
   * Median imputation for numerical variables
2. **Data Cleaning**:
   * Removal or correction of negative values
3. **Feature Engineering**:
   * Label encoding for categorical data
   * Standardization using StandardScaler
   * Outlier detection and treatment
4. **Train-Test Split**:
   * 80% training and 20% testing
   * Random state fixed for reproducibility

**4. Model Implementation**

**1. PyCaret AutoML**:

* Automatically trained and compared multiple models
* Helped select top-performing regressors

**2. LightGBM Regressor**:

* Parameters:
  + n\_estimators: 100
  + learning\_rate: 0.1
  + num\_leaves: 31
  + objective: regression
  + random\_state: 42

**5. Model Evaluation**

| **Metric** | **Value** |
| --- | --- |
| MSE | 3615.27 |
| RMSE | 60.13 |
| MAE | 30.84 |
| R² Score | 0.9016 |

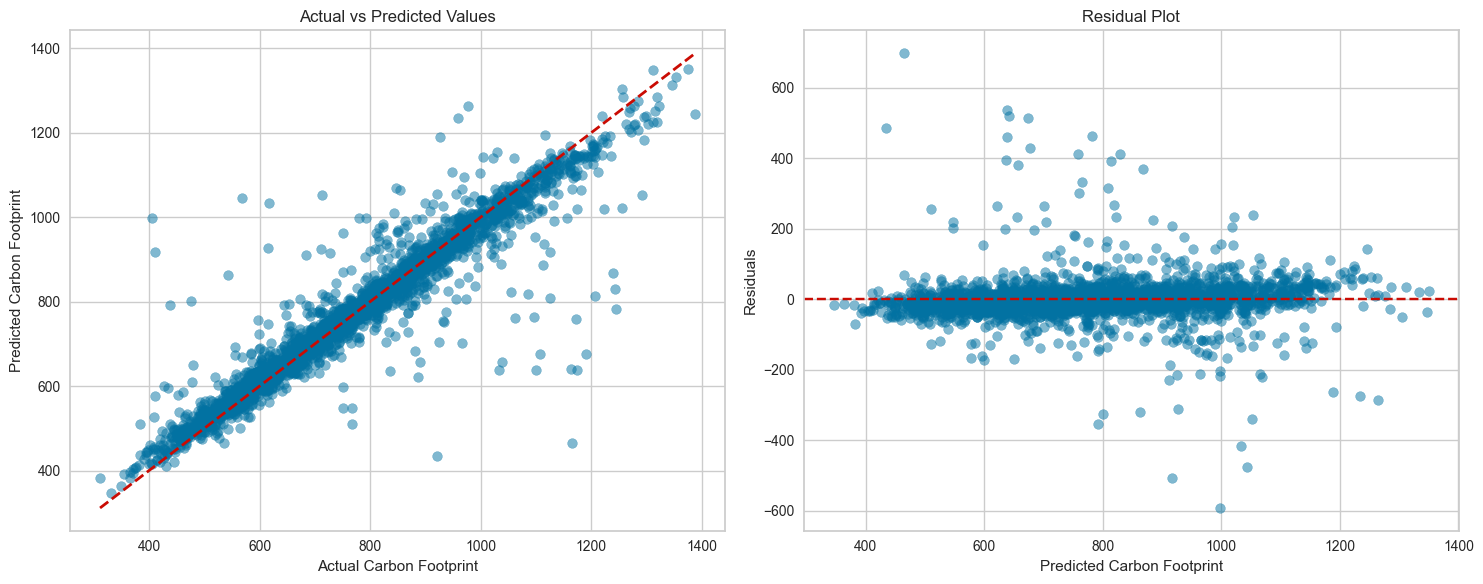
The R² score of 0.9016 indicates that 90.16% of the variance in carbon footprint can be explained by the model.

**6. Visualizations**

**Feature Importance Plot**

* Top features:
  + Meat consumption
  + Vehicle miles
  + Natural gas and electricity consumption
  + House area and insulation

**Actual vs Predicted Values and Residual Plot**



**7. Residual Statistics**

* **Mean of Residuals**: -0.61
* **Standard Deviation**: 60.13
* **Skewness**: 1.59

Residuals are slightly right-skewed but centered around zero, suggesting reasonable model calibration.

**8. Conclusion & Insights**

* The model is effective at predicting carbon footprints based on lifestyle factors.
* Transportation and energy usage are critical contributors.
* The project demonstrates practical use of ML in environmental analysis.

**9. Future Work**

* Introduce seasonal patterns in features
* Apply hyperparameter optimization
* Explore ensemble and neural models
* Integrate more granular behavioral data
* Employ cross-validation for robust evaluation