Machine Learning - Regression

Sami Taider

March 25, 2025

1 Introduction

The goal of this project is to predict car emissions (CO_2) from various car attributes such as weight, fuel type, and consumption metrics. Machine learning techniques such as Decision Trees, Random Forests, and XGBoost are explored to model the relationship between car features and CO_2 emissions.

2 Data Overview

The dataset consists of car specifications with the following columns:

- `id`, `brand`, `model`, `car_class`, `range`, `fuel_type`, `hybrid`, `max_power`, `grbx_type_ratios`, `weight_min`, `weight_max`, `urb_cons`, `exturb_cons`, `overall_cons`, `co`, `hc`, `nox`, `hcnox`, `ptcl`, `co2`.

The target variable is the CO_2 emission (`co2` column).

3 Data Preprocessing

3.1 Step 1: Feature Selection and Cleaning

Initially, columns like `id`, `model`, `range`, and `hc` were identified as unnecessary for the model due to their limited relevance or missing values:

- The `id` column is a unique identifier and does not contribute to prediction.
- The `hc` column had too many missing values and was therefore dropped.
- The `model` and `range` columns were considered non-informative for the task at hand.

3.2 Step 2: Handling Missing Values

The missing values in the dataset were handled using two approaches:

- In the first iteration, missing values for numerical columns such as `nox`, `co`, `hcnox`, and `ptcl` were filled with the mean of the respective columns.
- In the second iteration, rows with missing values were dropped, resulting in a smaller but more complete dataset.

3.3 Step 3: Encoding Categorical Variables

Non-numerical columns such as `brand`, `car_class`, and `fuel_type` were considered for encoding. Initially, One-Hot Encoding was applied to transform categorical variables into numerical values. However, this was later abandoned in favor of dropping non-numerical columns entirely, as the performance of the model improved by reducing feature complexity.

3.4 Step 4: Splitting Data

The dataset was split into training and test sets using an 80/20 split. The target variable is `co2`, and the features were all other columns except for the target variable.

4 Model Training and Evaluation

4.1 Initial Model Selection

Several regression models were tested, including:

- `RandomForestRegressor`
- `XGBRegressor`
- `DecisionTreeRegressor`

The models were evaluated using Mean Absolute Error (MAE) as the primary metric. The results for each model are shown below:

- Decision Tree: MAE = 0.09598496936899942
- XGBoost: MAE = 0.27972621869179803
- Random Forest: MAE = 0.09001164066109175

From these results, the `DecisionTreeRegressor` provided the best performance.

4.2 Hyperparameter Tuning

After identifying the Decision Tree as the best model, hyperparameter tuning was performed using `GridSearchCV` to find the optimal configuration. The parameter grid included:

- `max_depth`: [5, 10, 15, None]
- `min_samples_split`: [2, 5, 10]
- `min_samples_leaf`: [1, 2, 5]

After performing grid search, the best model was selected and saved using `joblib` for later use. The best hyperparameters were identified, and the model was retrained with these optimal settings.

4.3 Model Evaluation

The best model was evaluated on the test set using MAE and relative error. The final results are:

- **MAE = 0.04884945500470767**
- **Relative Error = 0.02%**

5 Analysis of Results - Data Processing Impact

The initial approach of filling missing values with the mean led to a higher MAE. However, after dropping rows with missing values and non-informative columns like `id`, `model`, and `range` as well as presumably informative, but non-numerical columns like `fuel_type` and `car_class`, the model performance improved significantly. The Decision Tree model achieved an MAE of 0.04885 with a relative error of just 0.02%.

By removing irrelevant features, the model complexity was reduced, preventing overfitting and enhancing its ability to generalize to unseen data. Furthermore, eliminating categorical encoding and simplifying the feature set resulted in a more efficient model that focused on the most relevant variables, improving both accuracy and computational efficiency.