sustainability

MDPI

Article

Assessment of CO_2 Emissions for Light-Duty Vehicles Using Dynamic Perturbation Additive Regression Trees

Hang Thi Thanh Vu and Jeonghan Ko *

Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Suwon-si 16499, Republic of Korea
* Correspondence: jko@ajou.ac.kr; Tel.: +82-31-219-2335

Abstract: Effective predictive modeling is crucial for assessing and mitigating energy consumption and CO_2 emissions in light-duty vehicles (LDVs) throughout the whole value chain of an organization. This study enhances the modeling of LDV CO_2 emissions by developing novel approaches to analyzing vehicle feature datasets. New tree-based machine learning models are developed to increase the accuracy and interpretability in modeling the CO_2 emissions in LDVs. In particular, this study develops a new algorithm called dynamic perturbation additive regression trees (DPART). This new algorithm integrates dynamic perturbation within an iterative boosting framework. DPART progressively adjusts prediction values and explores various tree structures to improve predictive performance with reduced computation time. The effectiveness of the new ensemble-tree-based models is compared to that of other models for the vehicle emission data. The results demonstrate the new models' capability to significantly improve predicting accuracy and reliability compared to other models. The new models also enable identifying key vehicle features affecting emissions, and thus provide valuable insights into the complex relationships among vehicle features in the dataset.

Keywords: CO_2 emissions; emission assessment; predictive modeling; tree ensemble; light-duty vehicle; sustainable value chain; Scope 3 emissions

Citation: Vu, H.T.T.; Ko, J. Assessment of CO_2 Emissions for Light-Duty Vehicles Using Dynamic Perturbation Additive Regression Trees. *Sustainability* **2024**, *16*, 10335. https://doi.org/10.3390/su162310335

Academic Editor: Antonio Caggiano

Received: 4 November 2024
Revised: 18 November 2024
Accepted: 20 November 2024
Published: 26 November 2024

1. Introduction

Growing concerns about carbon dioxide (CO_2) emissions from light-duty vehicles (LDVs) such as cars, vans, and light trucks have increased the need to accurately assess these emissions for carbon footprint deduction. In 2022, the total CO_2 emissions from LDVs were around 3.5 billion tons, which is a significant portion of the total global emissions, accounting for approximately 5 percent [1]. These huge emissions should be reduced 6% annually until 2030 to achieve the Net-Zero emissions target [1]. LDV CO_2 emissions are also a significant portion of the greenhouse gas (GHG) emissions in the whole value chain of any organizations [2]. Regulatory agencies such as the Environmental Protection Agency (EPA) [3,4] and various national agencies within the European Union (EU) [5] are implementing progressively stricter monitoring protocols and lowering limits on the GHG emissions from LDVs. For instance, recently, the EU's 'Fit for 55' package [6] and the EPA's proposed 2027 emissions standards [7] are expected to play pivotal roles in reducing emissions from the transportation sector. Such efforts to reduce these emissions have driven a growing need for effective predictive models that can accurately evaluate emission levels based on various vehicle characteristics. Accurate emission modeling is also especially necessary in assessing the GHG emissions for Scope 3 reporting and reduction planning [2]. The Scope 3 emissions may account for a large portion of the GHG emissions of a corporate value chain due to the huge amount of transportation of goods within the value chain, but still are hard to evaluate accurately due to a variety of factors influencing vehicle emissions [2].

There is a need for advanced modeling approaches that can better handle the challenges in LDV emission modeling. Various advanced machine learning approaches have

been employed to enhance predictive modeling for emissions. Although previous studies have provided valuable insights into the factors influencing emissions [8], they have often relied on linear or less flexible modeling techniques that may not fully capture the underlying patterns in the data. Moreover, CO_2 emissions from LDVs are influenced by a complex interplay of diverse vehicle attributes and operational conditions. This complexity makes it difficult for conventional models to achieve high accuracy and consistency.

Developing models that balance accuracy, interpretability, and computational efficiency is essential to address the multifaceted nature of vehicle emissions. Regression trees provide a basic approach to modeling, offering simplicity and interpretability. Other tree-based methods, such as random forests and boosting, offer significant advantages in managing complex datasets and improving prediction accuracy. However, these methods can suffer from interpretability issues and high computational costs. Boosting can sometimes overfit if it is not carefully tuned, particularly with excessive model complexity.

To address these challenges, this study proposes a new algorithm called dynamic perturbation additive regression trees (DPART) to enhance predictive modeling of CO_2 emissions in LDVs. DPART employs a dynamic perturbation mechanism within an iterative boosting framework to continuously refine model performance. This approach allows DPART to progressively adjust predicted values and explore various tree structures to capture complex interactions in high-dimensional datasets. The algorithm also allows DPART to achieve comparable results with other tree-based models while significantly reducing computation time.

This study makes several contributions to the predictive modeling for LDV CO_2 emissions. First, it introduces DPART, a novel algorithm that leverages dynamic perturbation. Dynamic perturbation allows DPART to either fit a completely new tree to explore global patterns or perturb the previous tree to refine the current model's performance. DPART shows substantial predictive accuracy and model stability over traditional tree-based methods by fitting a collection of trees that capture complex patterns, refine performance, and avoid overfitting. Second, this paper demonstrates the efficiency of DPART. The results indicate that DPART achieves comparable prediction accuracy with significantly less computation time, compared to conventional models with larger iteration settings. The dynamic perturbation mechanism allows for incremental improvements, optimizing the model-building process and minimizing redundant computations. Third, the new prediction models quantify variable importance, enhancing the understanding of factors affecting emission levels. Overall, the findings show the potential of the new models to improve emission modeling and help establish effective strategies for assessing and reducing vehicle emissions in the whole supply chain [2,9]. By offering more accurate and robust CO_2 emission predictions, the new models overcome the limitations of conventional models [8].

This paper is structured as follows: Section 2 overviews the relevant literature. Section 3 presents dataset characteristics, tree-based prediction methods and evaluation metrics. Section 4 analyzes and discusses the results. The conclusions are given in Section 5.

2. Review of the State-of-the-Art Research

The challenge of accurately modeling CO_2 emissions from light-duty vehicles (LDVs) has received increasing attention due to its critical implications for environmental sustainability, regulatory compliance, and emission reporting [2,10,11]. The complexity of vehicle emissions arises from a myriad of factors, such as engine sizes, fuel types, driving conditions, and vehicle design. Traditional linear models, while simple and easily interpretable, often struggle to capture these intricate relationships fully [12]. Given these challenges, non-linear models may provide a more effective means of capturing the intricacies of emissions data.

Conventional non-linear models overcome the limitations of such linear models. These models have been known as more robust alternatives, offering enhanced accuracy by accommodating the complex interactions between variables [13]. Generalized additive models (GAMs) and other non-linear approaches have demonstrated the ability to capture

these interactions [14]. GAMs provide improved predictive performance but are less interpretable compared to linear models. GAMs are also restricted to additive structures, which can lead to missing important interactions between variables. These characteristics can limit their practical applicability in real-world scenarios [13,14].

Tree-based ensemble methods provide a flexible approach to dealing with the complexity in the datasets. Advancements in ensemble methods, such as random forests, bagging, and boosting, have addressed some of these limitations by combining multiple models to improve accuracy and robustness [15–17]. These methods can model complex relationships and interactions between features more effectively. Most ensemble methods were developed to mitigate overfitting. Ensemble methods outperform traditional linear models in predictive accuracy [18]. For instance, a bagged tree ensemble regression model shows better performance than linear models in predicting train body vibrations and reduces data requirements without additional monitoring equipment [19]. Random forests have been effectively applied to assess vehicle fuel efficiency, capturing non-linear interactions between features [20]. Gradient boosting has shown success in modeling traffic flow patterns with improved accuracy compared to linear regression [21]. Despite improvement, some of these methods can still struggle with bias [22].

The Bayesian additive regression trees (BART) method represents a significant advancement in tree-based modeling [23]. BART incorporates a Bayesian framework to improve model predictions and reduce bias. Although BART has been effective in various applications, its performance can fluctuate depending on configuration parameters and computational resources [24,25]. For instance, BART has been successfully used in healthcare, but it has shown sensitivity to the choice of priors and the complexity of the tree structure, which can affect its predictive performance [26]. In environmental modeling, BART's accuracy can vary with dataset size and the computational power available [27,28]. This sensitivity indicates that while BART is effective, further refinements are needed to improve efficiency and manage computational costs, rather than solely focusing on improving accuracy.

Despite these advancements, significant gaps remain in the effectiveness of current models to capture the characteristics and complexities of LDV emissions. Previous studies have demonstrated that sophisticated models can improve prediction accuracy but often sacrifice interpretability or require extensive computational resources. This study aims to fill these research gaps in the existing literature.

3. Data and Methodology

This section presents the data characteristics and tree-based models used for predicting vehicle CO_2 emissions. It also introduces a new algorithm for prediction models and discusses the evaluation metrics of their performance.

3.1. Characteristics of the Emission Data

The data utilized in this study are comprehensive, with a wide range of emission-related data and vehicle features. The dataset was collected from the open data portal of the Canadian government [29]. The dataset spans the years 2014–2023 and includes 10,233 cases with CO_2 emissions, fuel consumption, and key specifications for various LDVs. CO_2 emissions range from 94 to 593 g/km. Vehicle attributes include engine sizes (engine displacement or cylinder volume), the number of cylinders in an engine, the number of gear steps, and fuel types. Engine sizes span from 0.9 to 6.8 L, engines can have 3 to 12 cylinders, and gear steps can be up to 10. Fuel types are categorized as diesel (type D), ethanol E85 (type E), regular gasoline (type X), and premium gasoline (type Z). These diverse feature ranges in the data and a sufficient number of data points for each category support a thorough analysis.

The dataset was further refined to improve analysis quality. The summary statistics of the CO_2 emission data are shown in Table 1 with categorical features not shown for display simplicity. In order not to distort the analysis, a few outliers were removed from

consideration: engine sizes over 8 L and cylinder counts of more than 16 for a small number of sports cars. Moreover, fuel consumption values were excluded in analysis, because CO_2 emissions are values calculated directly from fuel consumption values, as noted in [8]. A detailed description, preprocessing, and exploratory analysis of the dataset are available in other studies [8].

Table 1. Summary statistics of the dataset for CO_2 emission prediction.

	Gear Step	Engine Size	Cylinder Count	CO_2 Emissions
Min.	0	0.9	3	94
Median	6	3	6	248
Max.	10	6.8	12	593
Class	numeric	numeric	integer	numeric

The effects of LVD fuel types on CO_2 emissions are illustrated in Figure 1 in terms of the average and variability. The LDVs using ethanol E85 (type E) have higher average CO_2 emissions compared to those using the other fuel types. The LDVs with fuel type E release on average 277 g/km of CO_2, which exceeds 248 g/km for diesel. The broader range of emissions for ethanol E85 suggests inconsistencies in environmental performance compared to diesel. The LDVs with regular gasoline (type X) have an average emission of 236 g/km but show a wide range of emission levels. The LDVs with fuel type Z, though similar in median emissions to diesel, show the widest range of CO_2 emissions.

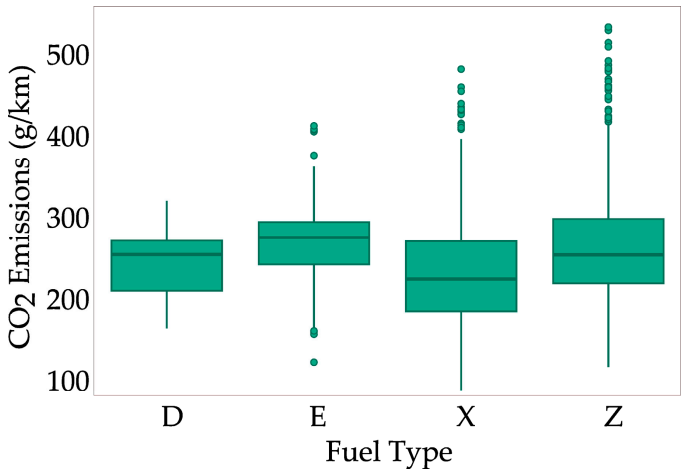

Figure 1. Average CO_2 emissions by fuel type.

LDV CO_2 emissions show an upward trend with engine size and cylinder count variables across different numbers of gear steps, as shown in Figure 2. The linear regression lines illustrate the association between CO_2 emissions and the two key vehicle attributes. The linear regression lines for each gear group based on engine size and cylinder count variables exhibit varying accuracy, with R-squared (R^2) values ranging from 0.24 to 0.93 for the engine size variable and 0.25 to 0.85 for cylinder count. There is an upward trend for each gear step. Figure 2a shows how CO_2 emissions vary with engine sizes across different gears, revealing a tendency for larger engines to generally emit more CO_2. Figure 2b shows the effect of the number of cylinders on CO_2 emissions with a similar pattern. The vehicles with more cylinders tend to emit more CO_2. These plots demonstrate the significant impact of both the engine size and cylinder count variables on emissions.

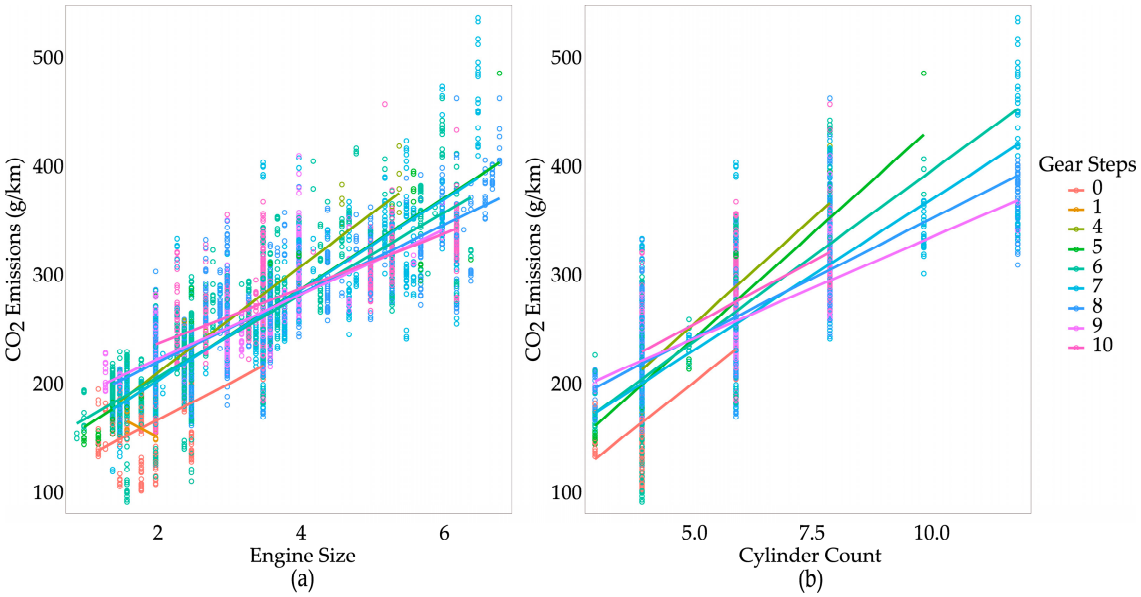

Figure 2. Relationship between CO_2 emissions and (**a**) engine size and (**b**) cylinder counts, grouped by the number of gear steps.

3.2. Tree-Based Methods

3.2.1. Existing Tree-Based Methods

The regression tree or decision tree (RT) method is a simple and straightforward approach for modeling relationships in data. A regression tree divides the predictor space into G distinct regions. In region R_g, the predicted value is the average response value ($\hat{y}_{R_g}$) of the training observations. These regions are created to minimize an error measure such as the residual sum of squares (RSS):

$$\text{RSS} = \sum_{g=1}^{G} \sum_{i \in R_g} \left(y_i - \hat{y}_{R_g} \right)^2 \quad (1)$$

Regression trees offer simplicity and interpretability but often do not perform well compared to other supervised learning techniques.

To enhance prediction accuracy and avoid bias, ensemble methods are widely used. Such methods include random forests, bagging, Bayesian additive regression trees (BART), and boosting. These methods comprise generating many trees known as weak learners and consolidating their predictions into a unified outcome.

In this paper, BART is also employed to compare prediction accuracy with the other methods. BART is a flexible, non-parametric Bayesian model that constructs an ensemble of trees to capture complex relationships in the data. The output is the sum-of-trees model, which is an additive model with multivariate components. Each tree in the ensemble contributes to the final combined prediction, which is more accurate than an individual prediction by a tree.

The construction of the trees in BART is designed to ensure comprehensive pattern capture. Each tree in BART is built randomly and tries to capture the patterns that the current set of trees has not yet captured. BART builds trees iteratively using the original data. In addition, each tree undergoes perturbation to ensure a more thorough exploration of the model space and prevent local optima. This approach prevents the model from being stuck in suboptimal solutions and improves generalization ability to new data.

BART's Bayesian framework provides a probabilistic interpretation of the model. It uses a Bayesian framework to estimate the tree structures and the terminal node parameters, allowing for a probabilistic interpretation of the model and its predictions. This is achieved through a Markov chain Monte Carlo (MCMC) sampler, which samples the model parame-

ters and trees from the posterior distribution [23]. MCMC enables the exploration of the complex, high-dimensional space of possible tree models, providing a way to approximate the posterior distribution and make robust inferences.

By averaging over multiple trees and leveraging the Bayesian framework, BART mitigates overfitting and improves the robustness of the predictions. This approach is particularly effective in scenarios with complex data patterns and interactions.

3.2.2. Dynamic Perturbation Additive Regression Trees

Motivated by the idea of BART, the dynamic perturbation additive regression trees (DPART) method is proposed to further improve tree models applied to the LDV dataset. BART constructs each tree in a random manner to capture residual signals by drawing new trees from a posterior distribution using MCMC sampling. Instead, DAPRT fits either a fresh new tree or the perturbed version of the previous tree on a modified version of the original training dataset.

DPART is a novel machine learning algorithm that enhances predictive performance by integrating a dynamic perturbation mechanism within an iterative boosting framework. DPART begins with an initialization phase where each tree's prediction is set, followed by the computation of an initial combined prediction. In the iterative process, the residuals are calculated based on the gap between the observed values and the current predictions. At each iteration, new trees are fitted to these residuals, with a probability-based mechanism to perturb the previous iteration's tree, thus exploring alternative tree structures. This perturbation, handled by a subroutine, involves randomly adjusting prediction values at the terminal nodes of trees, helping the model to avoid local optima and improve its generalization capability. The iterative process continues until the specified number of iterations is performed. The initial "warm-up" iterations are discarded, and the remaining predictions are averaged to produce the final output.

DPART consists of two parts: Algorithms 1 and 2. Algorithm 1 is the main procedure of DPART, managing the initialization, residual computation, and tree-fitting steps. Algorithm 2 is embedded within this process; Algorithm 2 is a subroutine responsible for introducing dynamic perturbations at the terminal nodes of trees.

Algorithm 1 shows the overall structure of DPART. The DPART algorithm begins by initializing each tree's prediction and computing an initial combined prediction. In this initial iteration ($m = 1$, where m is the variable denoting iteration number), the combined prediction is given by $\hat{f}^1(x) = \frac{1}{n}\sum_{i=1}^{n} y_i$ because all the T trees are initially set with a single root node, with $\hat{f}_t^1(x) = \frac{1}{nT}\sum_{i=1}^{n} y_i$.

In the remaining iterative process ($m > 1$), the partial residual is computed and used to update each tree, one at a time. The partial residual for each data point, r_i, is calculated as the difference between the actual target value y_i and the sum of $\hat{f}_t$ from all trees in the previous iteration, except for the tree currently being updated. This can be mathematically expressed as

$$r_i = y_i - \left(\sum_{t=1}^{T} \hat{f}_t^{m-1}(x_i) - \hat{f}_t^{m-1}(x_i)\right). \quad (2)$$

The calculated residuals are then used to adjust the corresponding tree $\hat{f}_t^m$, with the focus on refining the tree structure to improve the fit to the residuals and reduce the overall error. This process is repeated iteratively for each tree to gradually minimize residuals and improve prediction accuracy. Depending on a specified perturbation probability, either a fresh new tree or the perturbed version of the previous tree $\hat{f}_t^{m-1}$ is used to fit to the residual. Perturbing the tree allows exploration of alternative tree structures that may improve the fit to partial residual. After each iteration, the combined predictions are updated. Finally, the post-processing step computes the mean prediction after discarding a specified number of initial samples (warm-up).

Algorithm 1: The main procedure of DPART.

Input : $X_{train}, y_i, T, M, B, P$
Output : Final prediction function $\hat{f}(x)$
// Initialization: Initialize each tree's prediction
for t = 1 to T **do**
    $\hat{f}_t^1(x) = \left(\frac{1}{nT}\right)\text{sum}(y_i \textbf{ for } i = 1 \textbf{ to } n)$
end for
// Compute the initial combined prediction
$\hat{f}^1(x) = \text{sum}(\hat{f}_t^1(x) \textbf{ for } t = 1 \textbf{ to } T)$
end for
// Iterative Process
for m = 1 to M **do**
    for t = 1 to T **do**
        // Compute partial residual
        $r_i = y_i - (\text{sum}(\hat{f}_t^{m-1}(x_i) \textbf{for } t = 1 \textbf{ to } T) - \hat{f}_t^{m-1}(x_i))$
        // Fit a new tree with specified perturbing probability P
        if $m > 1$ **and** random () < P **then**
            $\hat{f}_t^m(x) = PerT(\hat{f}_t^{m-1}, X_{train}, r_i)$
        else
            $\hat{f}_t^m(x) = NewT(X_{train}, r_i)$
        end if
    // Compute the updated combined prediction
    $\hat{f}^m(x) = \text{sum}(\hat{f}_t^m(x) \textbf{ for } t = 1 \textbf{ to } T)$
    end for
end for
// Post-processing: Compute the mean prediction after B warm-up samples
$\hat{f}(x) = \frac{1}{M-B} sum(\hat{f}^m(x) \textbf{ for } m = B + 1 \textbf{ to } M)$

The subroutine for Algorithm 1, node perturbation algorithm, is shown in Algorithm 2. Algorithm 2 improves model performance by dynamic adjustment. As a component of DPART, it introduces randomness by perturbing prediction values at the terminal nodes of trees, evaluating the fit of the perturbed trees, and retaining the best-performing trees. This dynamic adjustment allows the model to adapt and improve over successive iterations, resulting in more accurate predictions. By leveraging perturbation and additive approaches, DPART aims to enhance the performance of tree-based models for the LDV dataset.

Algorithm 2: Node perturbation procedure of DPART.

Function $PerT(\hat{f}_t^{m-1}, X_{train}, r_i)$
    Input : Previous tree $\hat{f}_t^{m-1}$, X_{train}, residuals r_i
    Output : Best perturbed tree T_{best}
    for j = 1 **to** J **do**
        $T_j = \hat{f}_t^{m-1}$
        // Randomly perturb one terminal node
        $k = Random_{Uniform}(nodes)$
        $p = Random_{Uniform}(-\beta, \beta)$
        $T_j[k, \text{y}_{\text{val}}] = T_j[k, \text{y}_{\text{val}}] \times (1 + p)$
        // Evaluate the fit of the perturbed tree
        $\hat{r}_i = Predict(T_j, X_{train})$
        $current_{fit} = \text{sum}((r_i - \hat{r}_i)^2 \textbf{for } i = 1 \textbf{ to } n)$
        if $current_{fit} < best_{fit}$ **then**
            $best_{fit} = current_{fit}$
            $T_{best} = T_j$
        end if
    end for
    return T_{best}

The process in Algorithm 2 begins with function $PerT()$, which modifies the prediction at node k to introduce variability into the tree's structure. The inputs are the previous iteration's tree, training data, and residuals. The function outputs the best perturbed tree. In each iteration of $j = 1, \ldots, J$, where J denotes the number of tree variations considered, the algorithm creates a copy of the previous tree. A terminal node k is randomly selected from the tree, and its prediction value is perturbed by a random factor p, which is randomly chosen from a specified range $(-\beta,\ \beta)$.

After perturbing the tree, the algorithm evaluates its fit to the residuals on the training data and computes the sum of the squared differences between the actual and predicted residuals (RSS). This RSS score is used to compare the fit among J perturbed trees. If the current perturbed tree shows a better fit, indicated by a lower RSS $= \sum_{i=1}^{n} (r_i - \hat{r}_i)^2$, then the current tree becomes the new best tree T_{best}, and the best fit is updated accordingly.

Algorithm 2 allows flexible configuration of perturbation parameters in several ways. The number of nodes selected for perturbation can be varied based on the desired level of exploration, and the value of J can be adjusted to control how many perturbed versions of the trees are considered.

Depending on the specified perturbation probability value P, a fresh new tree can be fit to the current residual instead of being perturbed by the previous tree to better explore the global value. In this paper, the fresh new tree is created by function $NewT(X_{train}, r_i)$ using 'rpart' package in R for decision trees that model the relationship between the training data (X_{train}) and the residuals (r_i) [30]. By adjusting the perturbation probability, we can balance local and global optimization, enhancing the model's ability to improve predictive performance.

DPART enhances the performance by dynamically refining the tree structure perturbation. This approach progressively improves the model, leading to more accurate and robust predictions. The dynamic process ensures that the model continuously adapts and improves through successive iterations.

3.3. Evaluation Metrics

The performance of the models is assessed using two common metrics: root mean squared error (RMSE) and coefficient of determination (R^2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4}$$

Better model performance is indicated by lower RMSE and higher R^2 values. RMSE reflects the mean difference between the observed and predicted values.

To evaluate model consistency, Monte Carlo cross-validation (MCCV) is used to access the metrics' uncertainty with 50 sets of random train–test splits (70:30 ratio). The standard deviation (SD) of the metrics across these split sets indicates the uncertainty of the model's performance. Low SD values suggest consistent performance, while high SD values indicate variability.

4. Result Analysis and Discussion

This section details the results obtained from the tree-based prediction models and discusses practical implications of the models. The computational requirements for this study were modest. The models are run with generally brief execution times on a workstation with an Intel Xeon Silver 4210 CPU operating at 2.20 GHz and 96.0 GB of RAM.

4.1. Results from Decision Tree Models

Decision trees provide straightforward, interpretable models for predicting CO_2 emissions by breaking down complex relationships into simple, step-by-step decision rules. Decision trees can show the factors affecting CO_2 emissions and assist in making informed decisions regarding emission control.

The decision tree in Figure 3 shows a systematic approach to predicting CO_2 emissions based on key features, including engine sizes and gear steps. The tree consists of six internal and seven terminal nodes. Each terminal node represents a final predicted CO_2 emission level. Each node in the decision tree shows the predicted value for CO_2 emissions, along with the percentage of the observations that fall in that node. In Figure 3, the root node displays a value of 254, representing the average CO_2 emissions across all observations in the dataset. This node includes all (100%) data points before any splits occur. At the root, the model splits the data based on whether the engine size is less than 2.6 or not. The left branch, representing smaller engines (less than 2.6 L), covers 44% of the data and is further divided by additional splits, such as engine sizes below 1.9 and gear counts under three. The right branch with engines larger than 2.6 (56% of the data) shows further splits at engine sizes of 3.9 and 6.5, which narrows down the predictions to smaller subgroups of the data.

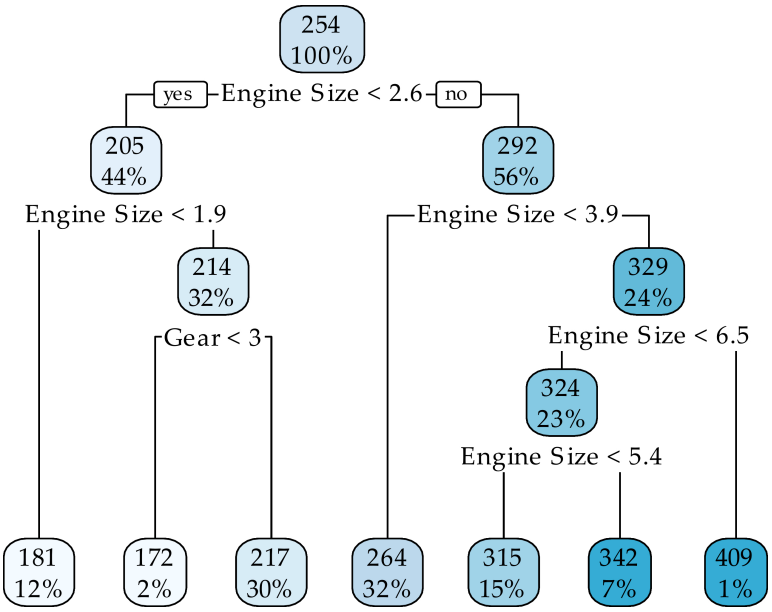

Figure 3. Decision tree for predicting CO_2 emissions.

In the results from the regression tree model, the engine size variable is the most significant factor in determining CO_2 emissions. Vehicles with smaller engines release lower emissions compared to those with larger engines. For vehicles with larger engines, the number of gears has a minimal impact on emissions. However, for vehicles with smaller engines, the number of gears does influence emissions, with those having more gears tending to release more CO_2.

Although decision trees offer clear and straightforward models that are easy to understand and visualize, they lack the predictive power and robustness of more advanced models. The regression tree model can produce a single tree that can vary with different data splits and the result may be misleading. For example, in Figure 3, the cylinder count variable does not appear because of the domination of the variable engine size and strong variable correlation.

4.2. Results from Dynamic Perturbation Additive Regression Trees

Dynamic perturbation additive regression trees (DPART) can significantly improve prediction accuracy. Figure 4 shows the performance of the DPART model on both the training and test sets over 1000 iterations. During the initial iterations, both training and test RMSE values highly fluctuate. However, after the warm-up period (B = 500), the RMSE values stabilize. The minimal difference between training and test RMSE indicates that DPART effectively mitigates overfitting and achieves convergence. The configuration of DPART in Figure 4 results in stable performance with an R^2 of 0.84 and RMSE of 23.9.

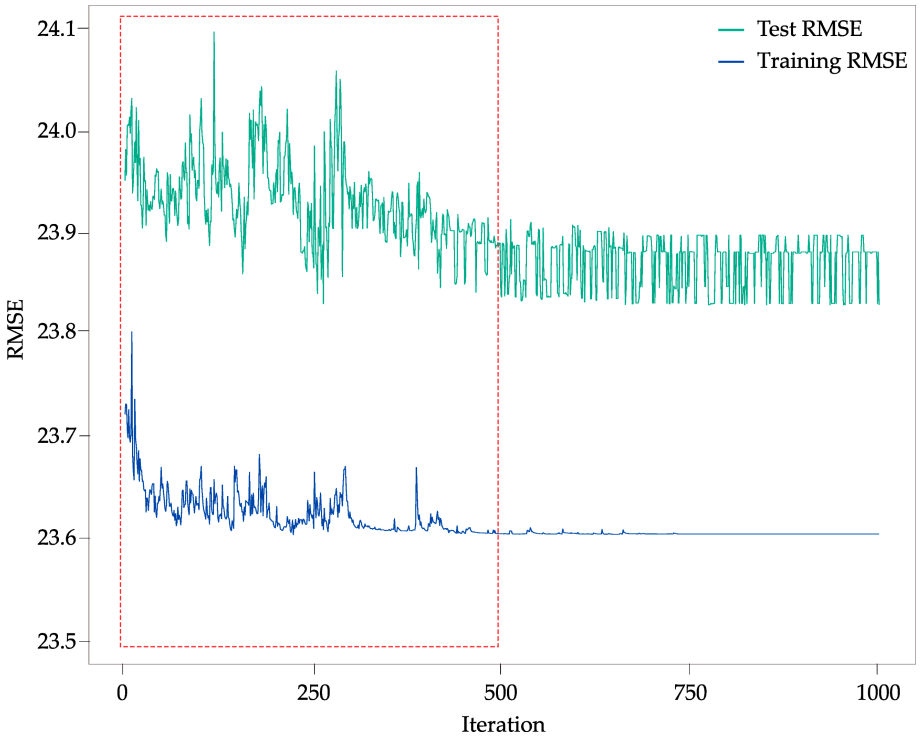

Figure 4. Training and test RMSE values over iterations for the DPART model. (Note: warm-up iterations are shown within the red box).

Various configurations were tested for DPART's capability to adjust setting for identifying an efficient balance between accuracy and computational time. One effective configuration consists of 5 trees, 30 iterations, and 15 warm-up iterations. This setup was determined by cross-validation results from the random forest model, which used 150 trees, aligning with 5 trees per iteration over 30 iterations. This streamlined configuration of DPART provides results comparable to those of larger iteration settings. The running time is reduced significantly from over 10 min in large DPART configurations to under 30 s in the smaller configurations.

The perturbation parameters were selected to maintain model stability while introducing meaningful variations. The perturbation probability is set to 0.5, ensuring a balanced chance of modifying the existing tree. For the $PerT()$ function, the previous tree is perturbed into 20 different versions by slightly adjusting the prediction value of a randomly chosen node. A random perturbation factor was chosen, empirically ranging from -0.01 to 0.01. To further validate the robustness of this configuration, a sensitivity analysis was conducted. Different perturbation ranges ($[-0.005, 0.005]$ and $[-0.02, 0.02]$) and probabilities (0.3 and 0.7) were tested. The results indicated minimal variations in predictive accuracy (R^2 fluctuating within ± 0.02) and computational efficiency, confirming that the chosen values provide a suitable balance between model performance and stability. In this paper, the $NewT()$ function is implemented using the regression tree package in R.

The variable importance obtained from the DPART model is shown in Figure 5. The variable importance shows the impact of each predictor on the model's output. Compared to regression trees, ensemble tree methods like DPART improve accuracy but at a cost of interpretability. However, some interpretability can still be accessed by computing the frequency of each variable's appearance across the ensemble of trees. The frequency analysis helps us to understand the importance of each variable. The engine size and cylinder count variables are identified as key factors in the model's predictions.

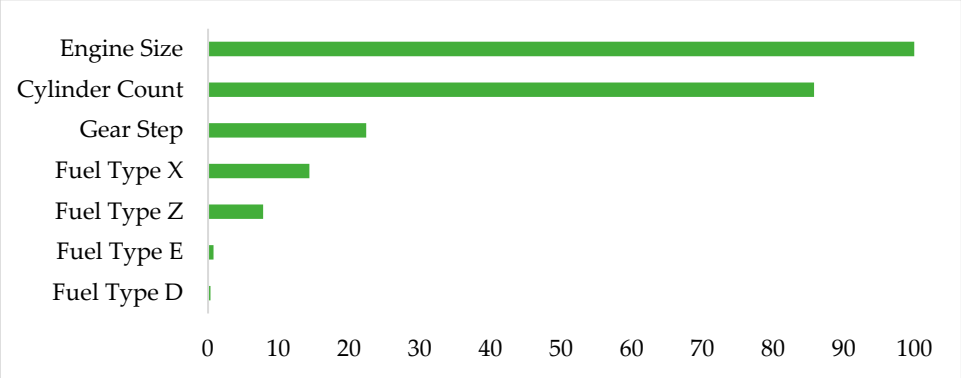

Figure 5. Variable importance resulting from DPART model.

4.3. Performance Comparison Among Tree-Based Models

To assess the effectiveness of different models compared to the proposed DPART model, performance metrics were analyzed for evaluating model prediction accuracy and consistency.

Table 2 and Figures 6 and 7 present the performance comparison between different models in terms of RMSE and R^2 values for their average and standard deviations (SDs) across 50 different sets of train–test splits in the MCCV. The models compared are a regression tree (RT) model, two BART models (BART–1 and –2), and a DPART model. BART–1 is a Bayesian additive regression tree model with relatively heavier computational requirement. BART–1 uses 200 trees, 1000 iterations, and 100 warm-up iterations [31] with default R package settings. BART–2 requires less computation with a configuration of 5 trees, 30 iterations, and 15 warm-up iterations. DPART use the same configuration for light computation requirement as BART–2.

As shown in Table 2 and Figures 6 and 7, DPART offers the best prediction accuracy compared to the other models showing mixed performance. The RT model shows a moderate fit with an average RMSE of 30.43 and an R^2 of 0.74. BART–1 outperforms RT. On the contrary, BART–2 performs worst. The DPART model with the same configuration as BART–2 shows the best performance with the lowest RMSE and highest R^2 values.

DPART also provides excellent prediction consistency along with high accuracy. The variation in RMSE of different tree-based models over the MCCV splits is shown in Table 2 and Figure 6. DPART shows minimal prediction errors and consistent performance with the lowest RMSE. The RT model shows moderate error levels with stable performance. BART–1 has low prediction errors with high consistency. On the other hand, BART–2 shows high prediction errors and performance variability with significantly higher RMSE and a large SD.

Table 2. Performance comparison between tree-based models.

	RMSE ± SD	**R^2 ± SD**
RT	30.43 ± 0.48	0.74 ± 0.0085
BART–1	24.94 ± 0.37	0.82 ± 0.0049
BART–2	35.34 ± 4.16	0.64 ± 0.0933
DPART	24.18 ± 0.37	0.84 ± 0.0052

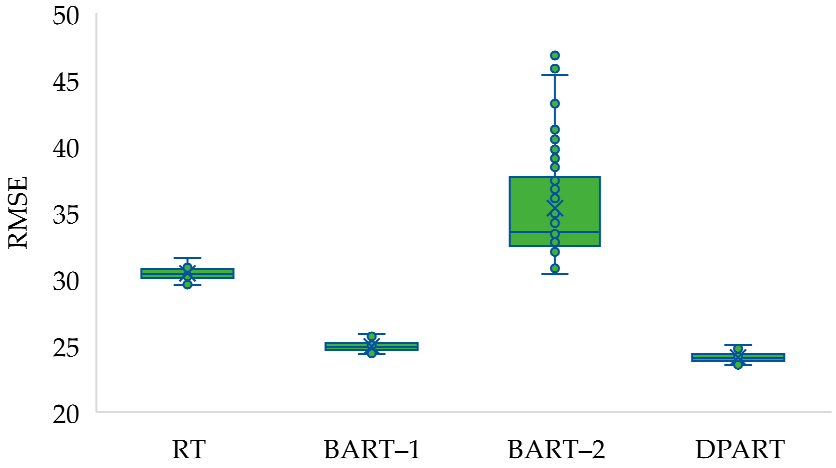

Figure 6. The variation in RMSE across 50 sets of MCCV train–test splits among different models.

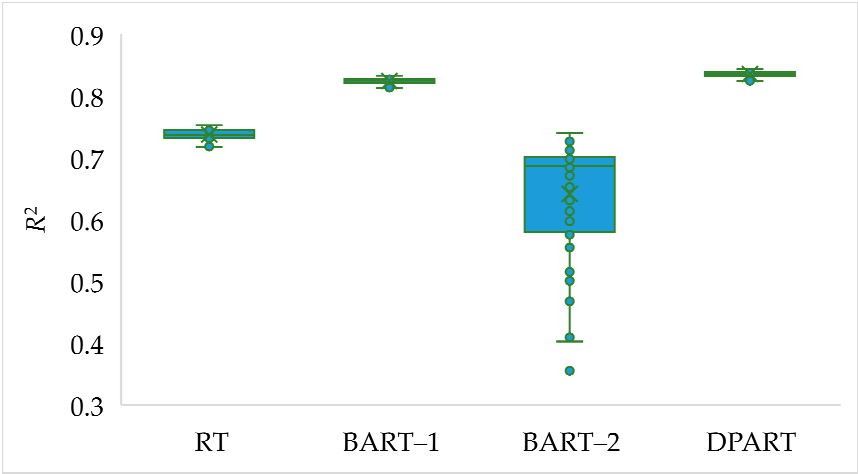

Figure 7. The variation of R^2 across 50 MCCV train–test splits among different models.

The performance measure R^2 shows similar characteristics to RMSE. The R^2 of different tree-based models over the MCCV splits is shown in Table 2 and Figure 7. DPART outperforms the other models in terms of accuracy and variability. The RT model shows a moderate fit and consistency with a moderate R^2 value. BART–1 achieves a higher R^2 with a low SD, indicating better accuracy and reliability. BART–2 shows lower accuracy and higher variability. The DPART model leads with the highest R^2 value at 0.84 and lowest SD at 0.0052, which indicate superior accuracy and consistent performance.

BART–2 consistently shows a large variation with the similar iterations and settings to DPART. This can be attributed to its reliance on MCMC sampling that inherently introduces stochasticity into the model estimation process. Despite MCMC being powerful in exploring complex posterior distributions, BART can also introduce variability in the results, particularly in configurations with fewer trees and iterations. With only 5 trees and 30 iterations, BART–2 may not have enough modeling power or convergence stability. BART–2 may struggle to consistently capture the underlying data structure. This leads to greater sensitivity to the initial conditions and the stochastic behavior of MCMC. As a result, the model may show higher variability across different runs and struggle to consistently converge to an optimal solution.

In this analysis, DPART performs well on this dataset due to its design. The RT model showed moderate performance, but it was outperformed by the more sophisticated BART–1 and DPART models. Simpler models like decision trees are less accurate compared to sophisticated models that can capture more intricate patterns in the data. Furthermore,

DPART is configured with a small number of trees and fewer iterations, yet it still achieves quite high accuracy consistently.

Table 3 shows the computational efficiency of DPART relative to other tree-based models. DPART achieves balanced performance by maintaining low computation times while preserving model accuracy. Unlike BART–1, which requires substantial computational time (70 s) due to its larger configuration of trees and iterations, DPART's setup is optimized for faster runtime without significantly compromising on predictive power. The simplest RT model has minimal computational requirements but lacks the predictive strength of ensemble methods like BART and DPART. BART–2 is also faster (0.5 s) but at the cost of very low accuracy due to its minimal settings. These results demonstrate that DPART offers an effective compromise.

Table 3. Computation time between tree-based models.

	Average Computation Time (Seconds)
RT	0.1
BART–1	70
BART–2	0.5
DPART	28

The comparison between the models in this study and a previous paper [8] reveals advancements in model performance. While the previous study reported a range of models with moderate accuracy, this paper demonstrates significant improvements. Results show that the DPART model and BART–1 exhibit superior performance, with lower RMSE and higher R^2 values compared to the best models from [8]. The prior study shows the limitations of linear models, which achieve RMSE values of approximately 31 g/km and R^2 values near 0.73, reflecting their inability to fully capture complex relationships in emissions data. For example, the R^2 values for the linear models are approximately 0.74 for the full feature set and 0.72 for a more interpretable model using just two variables (gear step and engine size). Non-linear methods, such as GAMs, demonstrated enhanced performance by surpassing the R^2 values of linear models by two to seven percentage points and reducing RMSE values by an average of 3 g/km. These findings show the advantage of non-linear approaches in addressing intricate patterns. However, these methods show only incremental improvements, with GAMs achieving an R^2 of around 0.77 using high-degree-of-freedom non-linear functions and four features. These findings reflect the strengths and limitations of traditional methods, where linear regression being valued for its simplicity and interpretability, but fails to capture complex patterns that non-linear models effectively identify.

The new models introduced in this study, particularly DPART, represent a substantial improvement. The new models generate R^2 values of approximately 0.84. This represents a substantial increase in prediction accuracy. This improvement indicates that the new models offer enhanced accuracy and better predictive capability, which reflects advances in modeling techniques and effectiveness in handling the dataset.

In addition to enhanced accuracy, DPART maintains a degree of interpretability despite the complexity typical of ensemble methods. Variable importance, computed by tracking the frequency of each predictor's presence across the ensemble of trees, shows that the engine size and cylinder count as influential variables in emission predictions. This finding aligns with previous studies using linear models and GAMs where the engine size and cylinder count variables were also identified as primary predictors.

5. Conclusions

This study presented new modeling of CO_2 emissions from LDVs by novel tree-based methods. A new algorithm was proposed to enhance predictive performance, interpretability, and computational efficiency in modeling the CO_2 emissions in LDVs. Comparative analysis revealed that the new models demonstrated substantial improvements in accuracy

and reliability compared to existing models. The results also showed that the new algorithm achieved comparable prediction accuracy with reduced computational time.

The new algorithm, DPART, achieves high prediction accuracy and consistency compared to existing models. DPART employs a dynamic perturbation mechanism within an iterative boosting framework, where each new tree is either a completely fresh model or a perturbed version of the previous one. This novel approach allows DPART to fit either a fresh new tree to explore global patterns or a perturbed version of the previous tree to refine the current model's performance. The iterative process of this method helps avoid local optima while progressively improving the model's accuracy. Unlike traditional regression trees, which typically rely on a single static tree structure, DPART builds a collection of trees, each contributing to a more robust and adaptable final model. This combination of global exploration and local enhancement enables DPART to capture complex patterns in the data while maintaining predictive consistency and avoiding overfitting. The dynamic process allows DPART to continuously refine its predictions, exploring various tree structures to capture complex interactions in datasets. The results demonstrate that DPART outperforms existing models in terms of accuracy and reliability.

In addition, the proposed DPART method achieves computational efficiency, significantly reducing the time required for model training and prediction. This efficiency is primarily driven by the iterative nature of the dynamic perturbation mechanism, which allows for incremental improvements rather than rebuilding the entire model from scratch. Furthermore, by fitting either fresh trees or perturbing existing ones, DPART minimizes redundant computations and optimizes the model-building process. During perturbation, the algorithm evaluates a set of perturbed trees and retains only the best performing ones, further improving computational efficiency. This approach allows DPART to selectively refine existing trees, accelerating the modeling process while maintaining high prediction accuracy. The study results illustrate DPART's capability to deliver rapid analysis while maintaining accuracy.

Variable importance analysis indicates that the engine size and cylinder count are significant predictors. These analysis results are consistent with the findings from earlier studies utilizing linear models and GAMs. This analysis is possible because DPART maintains a degree of interpretability alongside enhanced accuracy. Although ensemble methods like DPART often involve reduced interpretability, the quantification of variable importance through frequency analysis supports a clear understanding of emission factors. This interpretability, combined with DPART's predictive power, highlights its innovation compared to existing models.

The results demonstrate the key contributions of this research. First, the new prediction models in this study outperform conventional models in prediction accuracy and reliability. Second, with computational efficiency, the new methods achieve comparable prediction accuracy to that of conventional models. Third, the new models offer insights into the critical factors affecting CO_2 emissions. Therefore, this study can support reliable evaluation and informed decision-making on vehicle emission reduction related to Scope 3 assessment.

Although this study introduces new models and validates their effectiveness, it has limitations. The performance of DPART may vary with different datasets or perturbation settings, which could affect its generalizability. Furthermore, while DPART achieves high accuracy and computational efficiency, the interpretability of the model may be reduced as the complexity of the ensemble increases.

This study can be extended in several ways. The performance of the new models may be improved further. Future research could explore optimizing DPART's key parameters, such as random factors and tree sizes, to enhance performance and applicability across different datasets. Future research would include systematic parameter tuning, using techniques like grid search or Bayesian optimization, based on broader empirical results rather than partially relying on insights from previous models. Extending DPART to handle larger and more diverse datasets would offer a more comprehensive understanding of its capabilities and limitations. This could be achieved by leveraging parallel computing to dis-

tribute the computational workload and employing efficient data partitioning techniques, such as stratified sampling or clustering, to manage dataset complexity. Adaptive hyperparameter tuning and regularization strategies could be explored to optimize performance and prevent overfitting when dealing with heterogeneous data. Incorporating advanced feature engineering methods, such as automated feature selection or transformations, would further enhance DPART's ability to capture intricate patterns across diverse datasets.

Author Contributions: Conceptualization, H.T.T.V. and J.K.; methodology, H.T.T.V. and J.K.; software, H.T.T.V. and J.K.; validation, H.T.T.V. and J.K.; investigation, H.T.T.V. and J.K.; data curation, H.T.T.V.; writing—original draft preparation, H.T.T.V.; writing—review and editing, H.T.T.V. and J.K.; visualization, H.T.T.V.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (No. RS-2024-00400653), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1095569), and the Center for ESG at Ajou University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are accessible via following link: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64 (accessed on 10 August 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

Feature names	
CO_2 emissions	Carbon dioxide emissions (measured in grams per kilometer driven, g/km)
Cylinders	Number of cylinders per engine
Engine Size	Total engine displacement or cylinder volume (expressed in liters)
Fuel Type	D: diesel; E: ethanol E85; X: regular gasoline; Z: premium gasoline
Gear	Number of gear steps in a transmission (3–10)
Acronyms	
BART	Bayesian additive regression trees
DPART	Dynamic perturbation additive regression trees
EPA	Environmental Protection Agency
EU	European Union
GAMs	Generalized additive models
GHG	Greenhouse gas
LDVs	Light-duty vehicles
MCCV	Monte Carlo cross-validation
MCMC	Markov chain Monte Carlo
R^2	Coefficient of determination
RMSE	Root mean squared error (g/km of CO_2 emissions)
RSS	Residual sum of squares
RT	Regression tree
SD	Standard deviation
Mathematical Symbols	
B	Number of warm-up iterations
G	Number of regions ($g = 1, \ldots, G$)
J	Number of perturbation trees ($j = 1, \ldots, J$)
M	Number of iterations ($m = 1, \ldots, M$)
P	Perturbation probability
T	Number of trees ($t = 1, \ldots, T$)
k	Random terminal node
n	Count of observations

p	Random factor
x_i	ith input value
y_i	True values
$\hat{y}_i$	Prediction at x_i ($i = 1, \ldots, n$)
$\overline{y}$	Sample mean ($= \frac{1}{n}\sum_{i=1}^{n} y_i$)
$\hat{y}_{R_g}$	Average response in the gth box
y_{val}	Prediction values at the terminal nodes of trees
T_j	jth perturbed version of previous tree $\hat{f}_t^{m-1}$
T_{best}	Best perturbed tree
X_{train}	Training set
r_i	Partial residual for ith observation
$\hat{r}_i$	Predicted residuals for ith observation
$\hat{f}_t^m(x)$	Predicted value for CO_2 emissions at x for the tth tree in the mth iteration
$\hat{f}^m(x)$	Collection of prediction models at x in the mth iteration
$\hat{f}(x)$	Single prediction at x after B warm-up periods

References

1. International Energy Agency (IEA). Carbon Dioxide (CO_2) Emissions from Cars and Vans Worldwide from 2010 to 2022 (in Billion Metric Tons) [Graph]. In *Statista*. Available online: https://www.statista.com/statistics/1388092/carbon-dioxide-emissions-cars-vans-transport/ (accessed on 15 November 2024).
2. GHG Protocol. Corporate Value Chain (Scope 3) Standard. 1 May 2013. Available online: https://ghgprotocol.org/corporate-value-chain-scope-3-standard (accessed on 4 November 2024).
3. US EPA. Audit Protocols | US EPA. 16 November 2023. Available online: https://www.epa.gov/compliance/audit-protocols (accessed on 30 September 2024).
4. US EPA. Basic Information on Enforcement. 7 February 2024. Available online: https://www.epa.gov/enforcement/basic-information-enforcement (accessed on 30 September 2024).
5. European Climate Law. (n.d.) Climate Action. Available online: https://climate.ec.europa.eu/eu-action/european-climate-law_en (accessed on 30 September 2024).
6. European Council. Fit for 55. Available online: https://www.consilium.europa.eu/en/policies/fit-for-55/ (accessed on 11 November 2024).
7. U.S. Environmental Protection Agency. Final Rule: Multi-Pollutant Emissions Standards for Model Years 2027 and Later Light-Duty and Medium-Duty Vehicles. Available online: https://www.epa.gov/regulations-emissions-vehicles-and-engines/final-rule-multi-pollutant-emissions-standards-model (accessed on 11 November 2024).
8. Vu, H.T.T.; Ko, J. Effective Modeling of CO_2 Emissions for Light-Duty Vehicles: Linear and Non-Linear Models with Feature Selection. *Energies* **2024**, *17*, 1655. [CrossRef]
9. Vu, H.T.T.; Ko, J. Inventory Transshipment Considering Greenhouse Gas Emissions for Sustainable Cross-Filling in Cold Supply Chains. *Sustainability* **2023**, *15*, 7211. [CrossRef]
10. Tsiakmakis, S.; Fontaras, G.; Cubito, C.; Pavlovic, J.; Anagnostopoulos, K.; Ciuffo, B. *From NEDC to WLTP: Effect on the Type-Approval CO_2 Emissions of Light-Duty Vehicles*; Publications Office of the European Union: Luxembourg, 2017.
11. Commission Regulation (EU) 2017/1151. Official Journal of the European Union. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32023R0443 (accessed on 2 September 2024).
12. Murphy, K.P. *Probabilistic Machine Learning: Advanced Topics*; MIT Press: Cambridge, MA, USA, 2023.
13. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.
14. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: New York, NY, USA, 2023.
15. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
16. Sutton, C.D. Classification and Regression Trees, Bagging, and Boosting. *Handb. Stat.* **2005**, *24*, 303–329.
17. Ridgeway, G. Generalized Boosted Models: A Guide to the gbm Package. *Update* **2007**, *1.1*, 2007.
18. Seni, G.; Elder, J. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2010.
19. Peng, L.; Wang, J.; Liu, X.; Sun, Q.; Wei, L. A Novel Bagged Tree Ensemble Regression Method with Multiple Correlation Coefficients to Predict the Train Body Vibrations Using Rail Inspection Data. *Mech. Syst. Signal Process.* **2023**, *182*, 109543. [CrossRef]
20. Rathore, H.; Meena, H.K.; Jain, P. Prediction of EV Energy Consumption Using Random Forest and XGBoost. In Proceedings of the 2023 International Conference on Power Electronics and Energy (ICPEE); IEEE: New York, NY, USA, 2023; pp. 1–6.
21. Menguc, K.; Aydin, N.; Yilmaz, A. A Data-Driven Approach to Forecasting Traffic Speed Classes Using Extreme Gradient Boosting Algorithm and Graph Theory. *Phys. A Stat. Mech. Appl.* **2023**, *620*, 128738. [CrossRef]
22. Park, S.; Kim, C. Comparison of Tree-Based Ensemble Models for Regression. *Commun. Stat. Appl. Methods* **2022**, *29*, 561–589. [CrossRef]

23. Chipman, H.A.; George, E.I.; McCulloch, R.E. BART: Bayesian Additive Regression Trees. *Ann. Appl. Stat.* **2010**, *4*, 266–298. [CrossRef]
24. Hernández, B.; Mendez, A.; Figueroa, A.; Berrios, C.; Morales, J. Bayesian Additive Regression Trees Using Bayesian Model Averaging. *Stat. Comput.* **2018**, *28*, 869–890. [CrossRef] [PubMed]
25. Abu-Nimeh, S.; Wang, H.; Naderpour, M.; Poon, K. A Distributed Architecture for Phishing Detection Using Bayesian Additive Regression Trees. In Proceedings of the 2008 eCrime Researchers Summit; IEEE: New York, NY, USA, 2008; pp. 1–12.
26. Sparapani, R.; Spanbauer, C.; McCulloch, R. Nonparametric Competing Risks Analysis Using Bayesian Additive Regression Trees. *Stat. Methods Med. Res.* **2020**, *29*, 57–77. [CrossRef] [PubMed]
27. Wu, W.; Liu, X.; Li, Z.; Zhang, L.; Sun, W. Potential of Bayesian Additive Regression Trees for Predicting Daily Global and Diffuse Solar Radiation in Arid and Humid Areas. *Renew. Energy* **2021**, *177*, 148–163. [CrossRef]
28. Plant, E.; King, R.; Kath, J. Statistical Comparison of Additive Regression Tree Methods on Ecological Grassland Data. *Ecol. Inform.* **2021**, *61*, 101198. [CrossRef]
29. Fuel Consumption Ratings. Open Government Portal. Available online: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64 (accessed on 10 August 2024).
30. Therneau, T.; Atkinson, B.; Ripley, B. Package 'Rpart'. 2015. Available online: http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 September 2024).
31. Sparapani, R.; Spanbauer, C.; McCulloch, R. Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *J. Stat. Softw.* **2021**, *97*, 1–66. [CrossRef]