energies

MDPI

Article

Estimation for Reduction Potential Evaluation of CO_2 Emissions from Individual Private Passenger Cars Using Telematics

Masahiro Mae [1,*], Ziyang Wang [1], Shoma Nishimura [2] and Ryuji Matsuhashi [1]

[1] Department of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo 113-8656, Japan; wang@enesys.t.u-tokyo.ac.jp (Z.W.); matu@enesys.t.u-tokyo.ac.jp (R.M.)
[2] Department of Digital Business Design, Aioi Nissay Dowa Insurance Co., Ltd., Tokyo 150-8488, Japan; shoma-nishimura@aioinissaydowa.co.jp
* Correspondence: mae@enesys.t.u-tokyo.ac.jp

Abstract: CO_2 emissions from gas-powered cars have a large impact on global warming. The aim of this paper is to develop an accurate estimation method of CO_2 emissions from individual private passenger cars by using actual driving data obtained by telematics. CO_2 emissions from gas-powered cars vary depending on various factors such as car models and driving behavior. The developed approach uses actual monthly driving data from telematics and vehicle features based on drag force. Machine learning based on random forest regression enables better estimation performance of CO_2 emissions compared to conventional multiple linear regression. CO_2 emissions from individual private passenger cars in 24 car models are estimated by the machine learning model based on random forest regression using data from telematics, and the coefficient of determination for all 24 car models is $R^2 = 0.981$. The estimation performance for interpolation and extrapolation of car models is also evaluated, and it keeps enough estimation accuracy with slight performance degradation. The case study with actual telematics data is conducted to analyze the relationship between driving behavior and monthly CO_2 emissions in similar driving record conditions. The result shows the possibility of reducing CO_2 emissions by eco-driving. The accurate estimation of the reduced amount of CO_2 estimated by the machine learning model enables valuing it as carbon credits to motivate the eco-driving of individual drivers.

Keywords: CO_2 emissions; car fuel consumption; estimation; machine learning; driving data

check for updates

Academic Editor: Constantine D. Rakopoulos

Received: 14 November 2024
Revised: 11 December 2024
Accepted: 25 December 2024
Published: 27 December 2024

Citation: Mae, M.; Wang, Z.; Nishimura, S.; Matsuhashi, R. Estimation for Reduction Potential Evaluation of CO_2 Emissions from Individual Private Passenger Cars Using Telematics. *Energies* **2025**, *18*, 64. https://doi.org/10.3390/en18010064

1. Introduction

1.1. Research Background

The reduction in greenhouse gas emissions, such as carbon dioxide (CO_2), is necessary for a carbon-neutral society. In October 2020, Japan declared that it aims to achieve carbon neutrality by 2050 [1]. CO_2 emissions from the transportation sector account for about 20% of the total CO_2 emissions in Japan in 2020, and about half of them are emitted from private passenger cars [2]. Private passenger cars have the highest CO_2 emissions per unit of transportation compared to the other ways of travel used by passengers, such as airplanes, buses, and trains. Therefore, there is an increasing demand to reduce CO_2 emissions from private passenger cars.

In a macro analysis, CO_2 emissions from private passenger cars in Japan are calculated by a factor decomposition of CO_2 emissions with the fuel emission intensity, fuel efficiency, annual driving distance, and the number of vehicles. Comprehensive energy statistics, including an analysis of CO_2 emissions from private passenger cars in Japan, are used in planning energy policies, reporting Japan's energy supply and demand to the International Energy Agency, and reporting to the United Nations on energy-related CO_2 emissions in greenhouse gas emissions.

Although the macro analysis is important for such kinds of policy-making, a micro analysis of CO_2 emissions from individual private passenger cars is also important for reducing CO_2 emissions by changing the driving behaviors of individual drivers. For example, the Worldwide harmonized Light vehicles Test Cycle (WLTC) is commonly used to evaluate the fuel economy catalog value for each car model. These kinds of evaluation methods do not include the characteristics of each driver and are not suitable for encouraging eco-driving. The result in [3] shows that acceleration and velocity of the driving affect CO_2 emissions. The change in driving behavior influences CO_2 emission reduction, though the car model has a large impact on it [4]. The results in [5,6] show that the feedback of the driving behavior for eco-driving is effective for CO_2 emission reduction.

The accurate estimation of the CO_2 emissions from individual private passenger cars is the essential technology for feedback on the driving behavior of individual drivers for eco-driving. The CO_2 emissions from individual private passenger cars depend on various factors, such as the car model, car size, vehicle mass, engine type, driving behaviors, and season. The first principle modeling of the relationship between all of these features and CO_2 emissions from individual private passenger cars is too complicated, as such individual private passenger cars are driven by different drivers with different purposes. Therefore, it is necessary to select important features for estimating CO_2 emissions from individual private passenger cars and develop the estimation model with generalization performance from the limited data including these features. Accurate estimation performance of the CO_2 emissions from individual private passenger cars enables valuing the CO_2 emissions reduced by eco-driving as carbon credits, such as the J-Credit Scheme [7] in Japan, and it can be designed as incentives for eco-driving of individual private passenger cars.

1.2. Research Objective

With the spread of telematics [8], which can obtain driving data of private passenger cars, the estimation method of the CO_2 emissions from individual private passenger cars using actual driving data as big data can be developed. The aim of this paper is to develop an accurate estimation method of the CO_2 emissions from individual private passenger cars by using actual driving data obtained by telematics. Figure 1 shows the estimation model of the CO_2 emissions from individual private passenger cars. Vehicle features and driving data are used to estimate CO_2 emissions from individual private passenger cars.

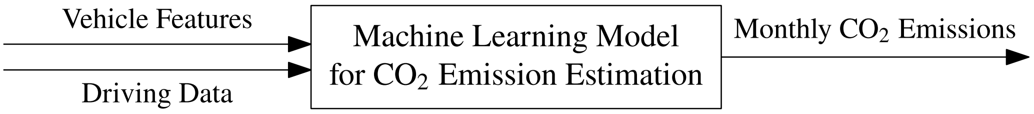

Figure 1. Model of monthly CO_2 emission estimation using driving data and vehicle features.

By estimating the CO_2 emissions from individual private passenger cars accurately, the amount of CO_2 reduction in eco-driving compared to the baseline can be valued as carbon credits. It has the potential to return value to individual drivers and to be used as an incentive for eco-driving and enables the reduction in the CO_2 emissions from individual private passenger cars.

1.3. Literature Review

In previous studies, several estimation models to estimate CO_2 emissions from individual private passenger cars are developed by using supervised learning such as linear regression, support vector machine, extra tree, random forest regression, multilayer perceptron, and deep learning. The nonlinear regression approaches enable higher estimation performance, and linear regression approaches can also benefit the feature analysis of the estimation model [9].

The utilization of data from telematics enables the more precise modeling and prediction of CO_2 emissions from cars. In macro aspects, it can be used for the development of smart cities [10] and the realization of intelligent transportation systems [11]. In micro aspects, it can be used for the spatiotemporal analysis of air pollution and climate change for urban design [12] and for real-time CO_2 emission estimation [13,14].

From these points of view, the literature review focuses on previous research about the CO_2 emission estimation of private passenger cars using actual driving data.

1.3.1. CO_2 Emission Estimation of Private Passenger Cars

There are several studies to analyze the features to estimate CO_2 emissions from individual private passenger cars. In [15], CO_2 emission factors are analyzed using hierarchical clustering. The insights of this approach can be used to select features in constructing a CO_2 emission estimation model. In [16], the driving behavior is classified into three groups, and the CO_2 emission analysis is conducted in both macro and micro aspects. In [17], the driving behavior is also classified into three groups respecting the difference between the Internal Combustion Vehicle (ICV) and Hybrid Electric Vehicle (HEV). These classifications of driving behavior and engine type are important to enhance the estimation performance of CO_2 emissions. In [18], real-time driving data are used to estimate vehicular fuel consumption on the highway using the artificial neural network, random forest regression, and reinforcement learning. In this analysis, the estimation performance is superior in random forest regression compared to the other two methods, but the estimation results are limited to the data of only four cars and are not directly applicable to other cars.

In this paper, it is necessary to build a generalized CO_2 estimation model that can estimate the CO_2 emissions from the car models that are not used in learning and that will be sold in the future. If the information of the car model number or the car name is directly used as a vehicle feature, it is not suitable for the generalization to estimate the CO_2 emissions for car models that are not used in learning. Therefore, it is necessary to select vehicle features that can deal with not only the limited car models used in learning but also those not used in learning and those that will be sold in the future.

1.3.2. CO_2 Emission Estimation Using Actual Driving Data

There are several estimation approaches of CO_2 emissions using actual driving data. In [19], a neural network is trained for each moving window to estimate CO_2 emissions during driving. This approach is suitable for the sequential estimation of time-series driving data but cannot be used to estimate the CO_2 emissions using driving data of each trip. In [14], the instantaneous CO_2 emission prediction method using multilayer perceptron regression is developed. The estimation method also cannot be applied to estimate the CO_2 emissions using driving data of each trip. In [20], the estimation accuracy of CO_2 emissions is compared using several machine learning methods. The comparison shows that the decision tree regression and the random forest regression achieve high accuracy in CO_2 emission estimation.

From these previous studies, it is important to select an appropriate machine learning method based on the size and characteristics of actual driving data. This paper deals with the estimation model that uses only monthly driving records with event-based acquisition because of the data format in telematics. The driving data used in this paper are not continuous time-series data but are nonlinear data with a count of monthly driving behaviors. Therefore, the CO_2 estimation method that can handle the nonlinear driving data is necessary.

1.4. Contributions

From the discussions in the literature review, the developed estimation model of the CO_2 emissions must fulfill the following requirements:

(R1) Estimation of CO_2 emissions that can be applied to various car models.
(R2) Utilization of monthly driving data with event-based driving behaviors.

Although important contributions have been made to develop the estimation methods of the CO_2 emission using machine learning with actual driving data, the estimation model of the CO_2 emission using monthly driving data generalized to the car model has not been developed. In this paper, the estimation method of the CO_2 emissions using machine learning is developed that uses vehicle features generalized to car models based on drag force and actual driving data. The driving data in this paper is the count of event-based driving behaviors for each month. In order to utilize such kinds of nonlinear driving data for CO_2 emission estimation, a machine learning model based on random forest regression is introduced to handle nonlinearity.

The contributions of this paper are as follows:

(C1) Generalization of the machine learning model to car models by using vehicle features based on drag force.
(C2) Estimation of CO_2 emissions using the machine learning model using actual monthly driving data for reduction potential evaluation.

The learning procedure for the machine learning model using big data takes a long time, so it cannot be updated frequently due to the limitation of computing resources in practice. When the estimation model of the CO_2 emission is used to value CO_2 emission reductions as carbon credits, the cost is incurred according to the amount of computing resources used by the business and the time required for calculation. Therefore, the estimation model that needs to be retrained every time a new car model is released is not economical from a business perspective. The generalized estimation model of the CO_2 emissions using vehicle features of the vehicle mass and the frontal area of private passenger cars can achieve sufficient estimation accuracy for car models not used in training, thereby reducing the frequency of retraining.

The remainder of this paper is as follows: In Section 2, the estimation method of the CO_2 emissions from individual private passenger cars using machine learning is described, constituting Contribution (C1). In Section 3, the estimation performance of the CO_2 emissions by the developed machine learning model using actual driving data obtained from telematics is evaluated, constituting Contribution (C2). In Section 4, conclusions are presented.

2. CO_2 Emission Estimation Model Using Machine Learning

In this section, an estimation model for monthly CO_2 emissions from individual private passenger cars using machine learning is developed. First, vehicle features are selected based on drag force, and it enables generalized estimation performance of the car models. Second, driving features are selected based on monthly driving data obtained from telematics, and it enables consideration of driving behaviors in the CO_2 emissions estimation. Finally, these selected features are constructed in a machine learning model with random forest regression to estimate monthly CO_2 emissions.

2.1. Vehicle Feature Selection Based on Drag Force

The vehicle dynamics based on drag force is used in vehicle feature analysis in order to construct the CO_2 emission estimation model generalized to car models. The drag force of the driving vehicle is modeled as follows [21]:

$$F = ma + mgC_r + 4v\frac{B_r}{r^2} + \frac{1}{2}Av^2\rho C_d \tag{1}$$

where each variable represents the physical quantity in Table 1. The drag force on the right-hand side consists of acceleration resistance, rolling resistance, viscous resistance, and air resistance, respectively.

Table 1. Symbols of vehicle dynamics model.

Symbol	Description	Unit
F	Driving Force	N
v	Velocity	m/s
a	Acceleration	m/s^2
g	Gravitational Acceleration	m/s^2
m	Vehicle Mass	kg
A	Frontal Area	m^2
ρ	Air Density	km/m^3
r	Wheel Radius	m
B_r	Viscous Damping Coefficient	Nms/rad
C_r	Rolling Resistance Constant	–
C_d	Air Resistance Coefficient	N/kN

The driving force is equal to the drag force, and it is related to fuel consumption that results in CO_2 emissions. Therefore, the variables in the drag force are used as vehicle features in the machine learning model. In this paper, the speed v and the acceleration a are not used as vehicle features because they are related to driving data. By using vehicle mass m and frontal area A as vehicle features, the machine learning model will learn parameters equivalent to gravitational acceleration g, air density ρ, wheel radius r, viscous damping coefficient B_r, rolling resistance constant C_r, and air resistance coefficient C_d. In addition, since engine type has a large influence on fuel efficiency [17], two engine types, Internal Combustion Vehicle (ICV) and Hybrid Electric Vehicle (HEV), are also treated as vehicle features.

2.2. Driving Feature Selection Based on Driving Data

The machine learning model is developed using monthly driving data obtained from telematics. In telematics, driving data such as vehicle speed, acceleration, accelerator, brake, shift lever, turn signal, headlights, and automatic braking can be acquired from sensors installed in the vehicle, and the obtained data are collected by uploading it to the internet server after each trip. The driving features in the driving data obtained from telematics are shown in Table 2. The features used in this paper consist of the monthly measurement period, driving time, driving distance, the number of speeding, sudden acceleration, sudden braking, and the safe driving score calculated from these driving behaviors. Each driving data are classified on expressways, national highways, and local roads. The driving data used in this paper are actual measurement data collected from the telematics installed in private passenger cars. The feature of the monthly measurement period enables considering seasonal effects on monthly CO_2 emissions due to such as using air conditioners. Note that since the research objective of this paper is an estimation of the monthly CO_2 emissions, the real-time driving data are not acquired because of too much data for the monthly timescale, and only monthly driving data are used for the

CO_2 emission estimation. The utilization of the real-time driving data from telematics for real-time CO_2 estimation can be seen in [14]. The data used in this paper do not contain geographic information such as temperature, humidity, and regional-dependent weather trends. Therefore, private passenger cars in extreme conditions that are far from the average of the training data should be estimated with other estimation models that are specialized to that region.

Table 2. Features obtained from telematics.

Feature	Unit
Target Year and Month	–
Start Date of Relevant Month	–
End Date of Relevant Month	–
Monthly Driving Time	s
Monthly Driving Distance	m
Speeding on Expressway	times
Sudden Acceleration on Expressway	times
Sudden Braking on Expressway	times
Driving Distance on Expressway	m
Speeding on National Highway	times
Sudden Acceleration on National Highway	times
Sudden Braking on National Highway	times
Driving Distance on National Highway	m
Speeding on Local Road	times
Sudden Acceleration on Local Road	times
Sudden Braking on Local Road	times
Driving Distance on Local Road	m
Safe Driving Score	%

2.3. *Learning Data of Monthly CO_2 Emission*

The monthly CO_2 emissions E_{CO_2} [$\mathrm{kg} - CO_2$] used for the machine learning model are calculated from the data of the monthly fuel consumption C_{gasoline} [L] as follows:

$$E_{CO_2} = 2.3 \times C_{\mathrm{gasoline}} \tag{2}$$

Since only the monthly fuel consumption data are available, the monthly CO_2 emissions are calculated indirectly from this formula. The monthly fuel consumption data are obtained by the measurement system that is independent of telematics. It is also actual measurement data with the same measurement period as that of telematics. By estimating monthly CO_2 emissions using monthly driving data, it is possible to provide feedback to encourage reducing CO_2 emissions through eco-driving.

2.4. *CO_2 Emissions Estimation Model Using Random Forest Regression*

The machine learning model for estimating CO_2 emissions from individual private passenger cars is constructed using vehicle features based on drag force and driving features based on driving data. In this paper, a multiple linear regression model is also used to compare with a random forest regression model in CO_2 emission estimation performance. Although other advanced machine learning methods, such as Support Vector Machine (SVM) regression and multilayer perceptron regression, can be used to build the CO_2 emission estimation model, these machine learning methods are difficult in feature analysis that is important for interpretability and explainability to encourage eco-driving. Compared to other regression methods, both multiple linear regression and random forest regression have the advantage of evaluating the contribution of each feature, and the feature analysis of the CO_2 emission estimation model helps to encourage eco-driving effectively. The driving data in Table 2 contain the number of counted driving behaviors. Therefore, the relationship between the features of these driving behaviors and CO_2 emissions is not linear, and nonlinear regression methods such as random forest regression are suitable for these driving features. Random forest regression requires a large amount of data for learning but has the advantage of high generalization performance and is suitable for estimating CO_2 emissions.

The learning procedure for random forest regression is shown in Figure 2. The learning procedure consists of the following four steps. In STEP 1, some data are sampled by bootstrap. In STEP 2, multiple different decision trees are trained. Note that each decision tree is overfitted to the sampled data at this step. In STEP 3, the estimation results of each decision tree are output. As a result, many different estimation results are obtained from many different decision trees. In STEP 4, the average of estimation results from each decision tree is calculated as a regression model. The overfitting can be prevented by ensembling multiple estimation results. Through these four steps, random forest regression estimates CO_2 emissions from ensemble learning of multiple bagged decision trees. For the learning process, after shuffling all the data, 80% is used as training data, and the remaining 20% is used as validation data. To prevent overfitting in random forest regression, five-fold cross-validation by grid search is applied as shown in Figure 3. The training data are divided into five parts, four of which are used for training and the other for validation. This process is repeated five times, and the average of the five training steps is used as the training result.

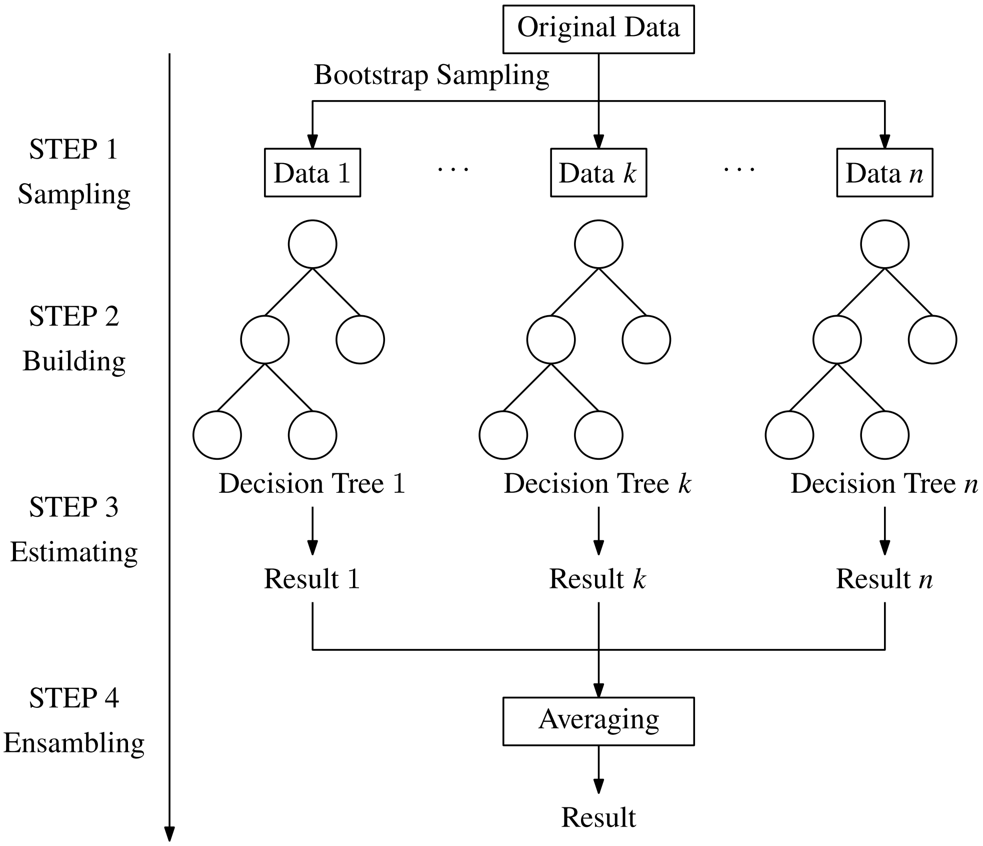

Figure 2. Procedure of random forest regression.

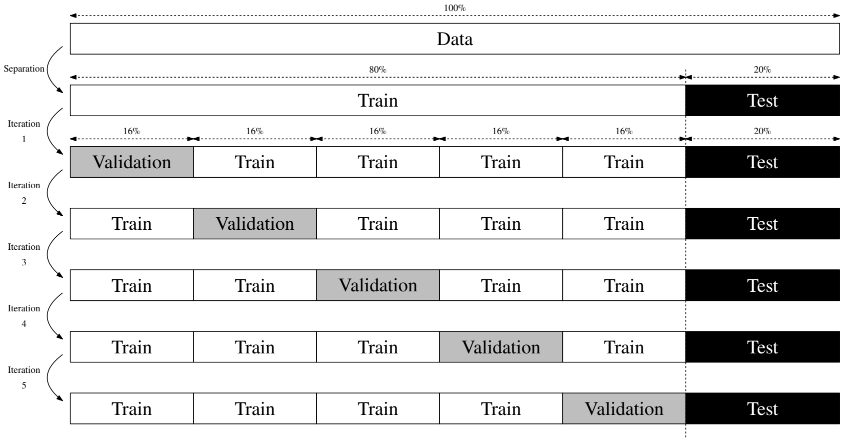

Figure 3. Five-fold cross-validation utilizing a grid search technique in random forest regression.

3. Model Evaluation with Actual Driving Data from Telematics

In this section, the performance of CO_2 emission estimation using actual driving data obtained from telematics is evaluated. First, the learning conditions of the CO_2 emission estimation model are described. Second, the learning results and estimation performance evaluation are presented in multiple linear regression and random forest regression. Third, feature analysis of the CO_2 emission estimation model is conducted. Fourth, the generalization performance of various car models is evaluated. Finally, the potential of eco-driving towards CO_2 emission reduction is discussed.

3.1. Learning Condition

For machine learning, the monthly driving data obtained from telematics in Table 2, the vehicle mass, frontal area, and engine type (ICV: 0, HEV: 1) of each vehicle are used as explanatory variables, and the data of monthly CO_2 emissions is used as the objective variable. Learning is conducted using normalized values for each feature. Details of the dataset used in this paper are shown in Table 3. Driving data are measured monthly, and the data are obtained from multiple users for the various car models. A total of n = 673,248 monthly driving data sets, excluding data with missing values, are used for learning. For machine learning, Scikit-learn [22] in Python 3 [23] is used. Learning is performed with the mean squared error as the loss function. In multiple linear regression, since the possibility of overfitting is low, five-fold cross-validation is not performed, and 80% of the data are used for training at once, and the remaining 20% of the data are used for evaluation. In random forest regression, the hyperparameters of the Scikit-learn function `RandomForestRegressor` are set empirically as follows, and training is performed on these combinations in a brute-force manner:

- `'n_estimators':[200,800,1400]`.
- `'max_features':['auto','sqrt']`.
- `'max_depth':[20,30,40]`.

Note that each hyperparameter has the following characteristics. `n_estimators` is the number of decision trees. The larger it is, the more expressive it is, but at the same time, the higher the computational cost. `max_features` is the number of features used by one decision tree for branching and determines the randomness of the decision tree. `max_depth` is the maximum depth of the decision tree. The larger it is, the more expressive it is, but the higher the possibility of overfitting. For machine learning calculations, the server with two Intel Xeon Gold 6258R CPUs, two NVIDIA Quadro RTX 8000 GPUs, and 768 GB of RAM is used.

Table 3. Specification of dataset.

No.	Count	%	m [kg]	A [m^2]	HEV = 1
1	64,457	9.56	2375	2.23	0
2	52,174	7.74	1465	1.73	1
3	41,897	6.21	1365	2.55	1
4	41,668	6.18	1865	1.85	0
5	41,375	6.14	1895	3.14	0
6	37,590	5.57	1645	2.56	1
7	35,398	5.25	1645	2.59	1
8	33,226	4.93	1405	2.53	1
9	29,749	4.41	1855	1.78	1
10	28,579	4.24	2045	2.63	1
11	28,404	4.21	1255	2.55	0
12	24,913	3.69	1715	2.79	1
13	24,295	3.60	1925	3.09	1
14	22,776	3.38	1385	2.81	0
15	22,748	3.37	2055	3.29	1
16	21,810	3.23	1965	2.29	1
17	21,300	3.16	1245	2.55	0
18	18,083	2.68	1685	2.96	1
19	16,876	2.50	2035	3.05	1
20	15,097	2.24	1525	2.81	1
21	14,581	2.16	1645	2.61	1
22	13,211	1.96	1605	2.56	0
23	13,166	1.95	2205	3.25	0
24	10,903	1.62	1945	2.59	1

3.2. Learning Result

Training is conducted in multiple linear regression and random forest regression to estimate monthly CO_2 emissions from individual private passenger cars. The estimation accuracy is evaluated using the coefficient of determination R^2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), defined as follows:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3}$$

$$\text{RMSE}(y,\hat{y}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{4}$$

$$\text{MAE}(y,\hat{y}) = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{5}$$

where the objective variable y, the estimated value $\hat{y}$, and the mean value $\bar{y}$.

The comparison of the monthly CO_2 emission estimation performance is shown in Table 4. In the estimation model after training in random forest regression, the best estimation results are obtained with the following combination of hyperparameters:

- `'n_estimators':1400`.
- `'max_features':'auto'`.
- `'max_depth':40`.

In this paper, a brute-force manner grid search is conducted for the hyperparameter tuning, and the estimation result with the best-case hyperparameters provides sufficient estimation performance as shown in Table 4. If the estimation performance is not sufficient for the application, other hyperparameter tuning methods such as Bayesian optimization can help to search for better hyperparameters efficiently.

Table 4. Total estimation performance of monthly CO_2 emissions and computation time of each machine learning method. Estimation models are trained by training data sets of 24 car models and validated by testing data sets of 24 car models.

Method	$\mathbf{R^2}$	**RMSE [kg−CO$_2$]**	**MAE [kg−CO$_2$]**	**Computation Time**
Multiple Linear Regression	0.874	30.8	19.9	<1 s
Random Forest Regression	0.981	12.0	7.55	>5 h

The coefficient of determination of multiple linear regression is $R^2 = 0.874$, which shows that bare minimum estimation performance can be achieved, and the calculation time is very short at less than 1 s. Since the model of multiple linear regression is simple, it has the advantages of evaluating the contribution of each feature, interpreting learning results, and preventing overfitting.

The coefficient of determination for random forest regression is $R^2 = 0.981$, which shows that a very high estimation performance can be achieved, but the calculation time is relatively long at more than 5 h. Note that random forest regression uses five-fold cross-validation in addition to ensemble learning to prevent overfitting and improve generalization performance.

Both multiple linear regression and random forest regression are known for high interpolation performance within the range of the variable space used for training, but the extrapolation performance is not as high as the interpolation performance. Therefore, it is better to conduct training using a relatively wide range of car models and driving data so that new car models and driving data not included in the dataset are included within the range of the variable space of the data used for training.

Furthermore, from social implementations to value CO_2 emission reductions as carbon credits, such as the J-Credit Scheme, there is a trade-off between the high coefficient of determination and the calculation time in multiple linear regression and random forest regression. An appropriate estimation method for the specific use of the estimated CO_2 emissions should be selected from the relative relationship between the estimated value, the estimation error, and the frequency of updating the learning data.

3.3. Feature Analysis

The regression coefficient and the absolute t-value of each feature in the multiple linear regression model and the feature importance in the random forest regression model are shown in Table 5. Note that in Table 5, each feature is listed in descending order of feature importance in the random forest regression model, and the regression coefficients in multiple linear regression are given for normalized explanatory variables. In multiple linear regression, the estimated equation and the t-value of each explanatory variable are given as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{6}$$

$$t_j = \frac{\beta_j}{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \times \frac{\left\{(X^{\mathsf{T}}X)^{-1}\right\}_{jj}}{n-p-1}}} \quad (j = 0, 1, \ldots, p) \tag{7}$$

where the objective variable is y, the explanatory variable is x, the regression coefficient is β, the number of data is n, the number of explanatory variables is p, and the matrix with rows as each data and columns as features is X.

In multiple linear regression, under the null hypothesis that the regression coefficient of each feature is 0, it is assumed that the test statistic follows a t-distribution, and the larger the absolute t-value, the more significantly the regression coefficient of each feature differs from 0. Therefore, using features with large absolute values of t-values in Table 5 is effective in the multiple linear regression model. Note that the number of data used in this paper is n = 673,248, which is a sufficiently large number, and even features with low correlation may have absolute values of t-values larger than 2, which is the 5% significance level.

Table 5. Feature analysis of estimation models. Coefficients and absolute t-value are shown for the multiple linear regression model. Feature importance is shown for the random forest regression model. Note that features are in descending order of feature importance.

Feature	Multiple Linear Regression		Random Forest Regression
	Coefficient	$\lvert t \rvert$-Value	Feature Importance
Monthly Driving Distance	4.40×10^{-1}	5.37×10^{1}	6.06×10^{-1}
Engine Type	-3.16×10^{-2}	1.29×10^{2}	1.61×10^{-1}
Vehicle Mass	4.85×10^{-2}	1.01×10^{2}	1.06×10^{-1}
Monthly Driving Time	-8.10×10^{-3}	3.12×10^{1}	9.61×10^{-2}
Driving Distance on Expressway	1.28×10^{-1}	1.07×10^{2}	4.49×10^{-3}
Driving Distance on National Highway	-2.19×10^{-2}	7.89×10^{1}	3.38×10^{-3}
Speeding on Expressway	1.48×10^{-1}	9.88×10^{0}	2.58×10^{-3}
Start Date of Relevant Month	6.60×10^{-3}	1.35×10^{0}	2.35×10^{-3}
Frontal Area	-1.78×10^{-3}	2.01×10^{1}	2.34×10^{-3}
End Date of Relevant Month	6.49×10^{-3}	1.94×10^{1}	2.32×10^{-3}
Driving Distance on Local Road	-5.19×10^{-4}	5.00×10^{1}	2.22×10^{-3}
Sudden Acceleration on Local Road	1.49×10^{-2}	1.04×10^{0}	1.87×10^{-3}
Sudden Braking on Local Road	2.46×10^{-2}	8.61×10^{0}	1.86×10^{-3}
Safe Driving Score	6.55×10^{-1}	1.56×10^{2}	1.72×10^{-3}
Sudden Braking on National Highway	-3.38×10^{-2}	5.88×10^{0}	1.24×10^{-3}
Sudden Acceleration on National Highway	3.73×10^{-2}	1.92×10^{0}	1.20×10^{-3}
Speeding on National Highway	6.31×10^{-2}	5.09×10^{0}	1.09×10^{-3}
Sudden Braking on Expressway	-1.60×10^{-2}	3.56×10^{0}	7.96×10^{-4}
Speeding on Local Road	2.22×10^{-2}	8.94×10^{-1}	4.14×10^{-4}
Target Year and Month	-2.55×10^{-3}	2.04×10^{0}	3.84×10^{-4}
Sudden Acceleration on Expressway	-2.97×10^{-3}	2.04×10^{0}	2.80×10^{-4}

In random forest regression, feature importance means how much Gini impurity, which is the normalized value of the reduction in the mean square of the prediction error multiplied by the weight of the number of data points, can be reduced by splitting the node using that feature. The explanatory variables with high feature importance can significantly reduce the Gini impurity by splitting the node using that feature. Note that feature importance does not represent the amount of change per unit amount.

Features with large absolute values of t-value and large feature importance are essentially important features in a CO_2 emission estimation model. In addition, when comparing the features of the driving behaviors and the safe driving score calculated from driving behaviors, those in random forest regression are around the same range. In the absolute t-value in multiple linear regression, the contribution of the safe driving score is sufficiently larger than that of the driving behaviors. The value of the driving behavior is 0 times in many cases and is a discrete value such as a single-digit integer value. On the other hand, the value of the safe driving score calculated from those driving behaviors is given as a value from 0 to 100, which is a relatively continuous value compared to the value of the driving behavior. Due to the difference in the continuity of the values of the driving behavior and the safe driving score and the possibility of multicollinearity, it is sufficient to use the safe driving score with a large absolute *t*-value as explanatory variables in the multiple linear regression model, rather than using the driving behaviors with a small absolute *t*-value. Note that the same features are used as explanatory variables in both random forest regression and multiple linear regression for comparison in this paper.

The contributions of the safe driving score and the driving behaviors in random forest regression are around the same because the random forest regression algorithm is nonlinear and has sufficient complexity to use the highly nonlinear driving behaviors as an explanatory variable. In addition, the feature importance of features other than monthly driving distance, engine type, vehicle mass, and monthly driving time, which have relatively high feature importance, is similar, and they have around the same contribution to the estimation accuracy. It can be confirmed from the relatively large feature importance values that random forest regression can take into account the influence of driving behaviors such as the number of speeding, sudden acceleration, and sudden braking, compared to multiple linear regression. Note that the number of explanatory variables has a trade-off between the complexity of the model and the training time.

Although CO_2 emissions can be reduced by the shorter monthly driving distance, which has the highest feature importance, there is a problem that deteriorates the convenience of private passenger cars for drivers. Therefore, it is beneficial to reduce CO_2 emissions by only improving driving behaviors and becoming eco-driving even if the car models, driving distance, and driving time are the same. For this reason, random forest regression, which can take into account the influence of driving behaviors in detail, is more effective for evaluating the CO_2 emission reduction due to eco-driving.

In conclusion, in feature analysis, when selecting features in a CO_2 emission estimation model, it is important to select features with a high contribution as explanatory variables according to the linearity and complexity of the regression model. It is expected that effective feedback for eco-driving can be achieved to reduce CO_2 emissions for individual users who drive private passenger cars by decomposing the features of CO_2 emissions and identifying importance.

3.4. Generalization Performance Evaluation

The generalization performance of the CO_2 emission estimation model is evaluated in multiple linear regression and random forest regression. In a previous study [24], the CO_2 emission is estimated by using random forest regression with training data including the feature of car models. Compared to the estimation approach in this paper, which uses vehicle mass and frontal area as vehicle features, the estimation approach in a previous study [24] is not generalized to car models because the car model information is directly used as a feature for training the estimation model. Although the developed model in this paper requires a long training time with a large amount of data, the trained estimation model is generalized to car models, and it does not require retraining for other car models frequently.

Figure 4 shows the distribution of vehicle mass and frontal area by engine type for the car models used in the driving data. There is no strong correlation between vehicle mass and frontal area in the dataset, and the possibility of multicollinearity between variables is low. In addition, the data of both Internal Combustion Vehicles and Hybrid Electric Vehicles is scattered across a wide range of variable space, and it leads to high generalization performance in the variable space. Generalization performance includes both interpolation performance, which estimates data within the range of the variable space of the training data, and extrapolation performance, which estimates data outside the range of the variable space of the training data. Multiple linear regression and random forest regression are known to have lower extrapolation performance than interpolation performance. From the distribution of vehicle mass and frontal area in Figure 4, interpolation performance is validated by using data from car model No. 12, and extrapolation performance is validated by using data from car model No. 2.

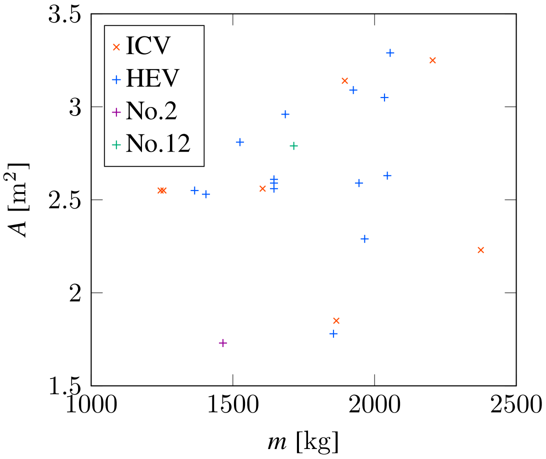

Figure 4. Distribution of vehicle mass and frontal area of 24 cars in driving data. ICV denotes Internal Combustion Vehicle and HEV denotes Hybrid Electric Vehicle, respectively.

The training on 23 car models, excluding the green cross of No. 12 in Figure 4, is conducted, and the CO_2 emission estimation performance of the trained model for car model No. 12 is evaluated to validate the interpolation performance. The results of the interpolation performance are shown in Table 6. The results show that a similar estimation accuracy is achieved compared to the estimation performance in Table 4 when all 24 car models are trained. The training on 23 car models, excluding the purple cross of No. 2 in Figure 4, is conducted, and the CO_2 emission estimation performance of the trained model for car model No. 2 is evaluated to validate the extrapolation performance. The results of the extrapolation performance are shown in Table 7.

Table 6. Interpolated estimation performance of monthly CO_2 emissions and computation time of each machine learning method. Estimation models are trained by data sets of 23 car models excluding No. 12 and validated by a data set of car model No. 12.

Method	**R^2**	**RMSE [kg−CO$_2$]**	**MAE [kg−CO$_2$]**	**Computation Time**
Multiple Linear Regression	0.846	23.4	14.9	<1 s
Random Forest Regression	0.970	10.3	6.81	>5 h

Table 7. Extrapolated estimation performance of monthly CO_2 emissions and computation time of each machine learning method. Estimation models are trained by data sets of 23 car models excluding No. 2 and validated by a data set of car model No. 2.

Method	**R^2**	**RMSE [kg−CO$_2$]**	**MAE [kg−CO$_2$]**	**Computation Time**
Multiple Linear Regression	0.763	24.8	19.7	<1 s
Random Forest Regression	0.965	9.60	6.59	>5 h

Comparing the interpolation performance in Table 6 and the extrapolation performance in Table 7, although the estimation accuracy is worse in the coefficient of determination when extrapolated, the estimation accuracy deterioration from interpolation to extrapolation is about 0.005 in random forest regression because the variable space of the trained data is relatively close and vehicle mass and frontal area are used in the developed approach. The result shows that a sufficiently high estimation accuracy is achieved even in the extrapolation performance validation using data from car model No. 2. In general, random forest regression is good at interpolation performance but not for extrapolation performance. Therefore, it is recommended to conduct training using a relatively wide range of car models and driving data so that the extrapolation point becomes relatively close to the range of the variable space of the training dataset. Note that the data used in this paper do not contain geographic information such as temperature, humidity, and regional-dependent weather trends. The generalization performance for the regions whose weather conditions are far from the average of the training data is not guaranteed. In such cases, CO_2 emissions should be estimated with other estimation models that are specialized to that region. In conclusion, the estimation performance is generalized to car models by using vehicle mass and frontal area as features even if the information of the car model to be estimated is not included in the training data.

3.5. Reduction Potential Analysis of CO_2 Emissions

Table 8 shows a case study for the reduction potential of CO_2 emissions. In Table 8, two sets of data for car model No. 2 are the same for the features with high feature importance, which are monthly driving distance, engine type, vehicle mass, monthly driving time, and driving distance on expressway. The comparison is made while excluding as much as possible influences other than driving behaviors, such as the number of times speeding and sudden acceleration and braking. The used data periods are from April to June, when the impact of air conditioners on fuel consumption is low. Note that the monthly driving distance is the total distance that includes the monthly driving distance on road categories other than expressway, national highway, and local road, such as narrow streets.

Table 8. Case study for comparison of monthly CO_2 emissions with different driving behaviors in data sets of car model No. 2.

Feature	Unit	Case A1	Case A2	Case B1	Case B2
Monthly Driving Distance	m	87,000	87,000	593,000	593,000
Engine Type	–	HEV	HEV	HEV	HEV
Vehicle Mass	kg	1465	1465	1465	1465
Monthly Driving Time	s	12,660	12,660	70,140	70,140
Driving Distance on Expressway	m	0	0	0	0
Driving Distance on National Highway	m	31,117	19,861	319,654	136,862
Speeding on Expressway	times	0	0	0	0
Start Date of Relevant Month	–	25 May 2022	1 May 2022	4 April 2022	7 May 2022
Frontal Area	m^2	1.73	1.73	1.73	1.73
End Date of Relevant Month	–	24 June 2022	31 May 2022	3 May 2022	6 June 2022
Driving Distance on Local Road	m	8692	20,508	20,546	270,401
Sudden Acceleration on Local Road	times	0	9	0	0
Sudden Braking on Local Road	times	1	4	1	5
Safe Driving Score	%	92	23	100	100
Sudden Braking on National Highway	times	1	4	2	3
Sudden Acceleration on National Highway	times	0	3	0	1
Speeding on National Highway	times	0	0	0	0
Sudden Braking on Expressway	times	0	0	0	0
Speeding on Local Road	times	0	0	0	0
Target Year and Month	–	May 2022	May 2022	April 2022	May 2022
Sudden Acceleration on Expressway	times	0	0	0	0
Monthly fuel consumption	L	4.273	5.379	20.079	21.007
Monthly CO_2 emission	$kg - CO_2$	9.828	12.37	46.182	48.316

Comparing Case A1 and Case A2, where the monthly driving distance is relatively short, Case A1, which is eco-driving with little sudden acceleration and braking, is about $2.5\,\mathrm{kg} - \mathrm{CO_2}$ lower in monthly CO_2 emissions than that of Case A2. Similarly, comparing Case B1 and Case B2, where the monthly driving distance is close to the average of the data set, Case B1, which is eco-driving with little sudden acceleration and braking, has about $2.1\,\mathrm{kg} - \mathrm{CO_2}$ lower monthly CO_2 emissions than that of Case B2. These comparisons show that estimating the CO_2 emissions from individual private passenger cars, taking into account the features of driving behaviors, can lead to effective feedback that encourages eco-driving for individual drivers. It is possible to reduce CO_2 emissions from individual private passenger cars by providing value return in carbon credits, such as the J-Credit scheme for CO_2 emissions through eco-driving. In the future, effective feedback that encourages eco-driving for individual users can be socially implemented by estimating the CO_2 emissions from individual private passenger cars, and the value of carbon credits for CO_2 emissions reduction through eco-driving can be returned to individual drivers.

4. Conclusions

CO_2 emissions from individual private passenger cars can be estimated by using machine learning with vehicle features and driving data. In this paper, the estimation method for CO_2 emissions from individual private passenger cars is developed by using the random forest regression with vehicle features based on drag force and driving data from the telematics. The developed estimation method uses the information of vehicle mass and frontal area as vehicle features instead of the information of the car model numbers or the car model names. CO_2 emissions estimation performance in 24 car models is $R^2 = 0.981$ in the coefficient of determination, and the estimation performance for interpolation and extrapolation of car models keeps enough estimation accuracy with slight performance degradation. The developed machine learning model realizes the generalized estimation performance for the inside and near the outside of the car models whose data are used in the training process. The generalized estimation performance is important to estimate CO_2 emissions for the car model that is not used in machine learning and for new car models that will be released in the future.

The reliable estimation method with generalization performance can be applied to the evaluation of CO_2 emissions using driving data from telematics to trade CO_2 emission reductions as carbon credits. The case study with actual telematics data is conducted to analyze the relationship between driving behavior and monthly CO_2 emissions in relatively the same driving record conditions. The result shows the possibility of reducing CO_2 emissions by eco-driving, and the accurate estimation of the reduced amount of CO_2 estimated by the machine learning model enables valuing it as carbon credits to motivate the eco-driving of individual drivers. CO_2 emission reductions by eco-driving of individual drivers can be valued by the developed estimation method, and the analysis of the improved eco-driving features can be applied to further feedback to individual drivers.

For the social implementation of the valuation as carbon credits, such as the J-Credit Scheme in Japan, the amount of CO_2 emissions is evaluated with the developed machine learning model using the monthly driving data of an individual private passenger car using telematics, and this evaluation stands on the accurate and generalized estimation performance of the CO_2 emission estimation model. The reduced amount of CO_2 emissions is valued by comparing it to the base case, such as the amount of CO_2 emissions before installing telematics or before the driving behavior is changed by the feedback in the previous month. The driving behavior change in the individual drivers in private passenger cars can contribute to reducing greenhouse gas emissions towards a carbon-neutral society.

Author Contributions: Conceptualization, S.N.; methodology, M.M. and Z.W.; software, M.M. and Z.W.; validation, M.M. and Z.W.; formal analysis, M.M. and Z.W.; investigation, M.M.; resources, S.N.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, M.M.; visualization, M.M.; supervision, R.M.; project administration, R.M.; funding acquisition, S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Aioi Nissay Dowa Insurance Co., Ltd.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Shoma Nishimura was employed by the company Aioi Nissay Dowa Insurance Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. This work was partly sponsored by Aioi Nissay Dowa Insurance Co., Ltd. The funding sponsor provided the telematics data (driving behavior data) essential for our research and proposed a hypothesis on the correlation between driving behavior data and monthly CO_2 emissions. While the sponsor facilitated the collection of data and offered a hypothesis for investigation, they had no role in the design of the study beyond these contributions; in the analysis or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ministry of Economy, Trade and Industry, Japan. Green Growth Strategy Through Achieving Carbon Neutrality in 2050. 2022. Available online: https://www.meti.go.jp/english/policy/energy_environment/global_warming/ggs2050/index.html (accessed on 28 October 2024).
2. Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism, Japan. Summary of the White Paper on Land, Infrastructure, Transport and Tourism in Japan, 2022. 2023. Available online: https://www.mlit.go.jp/en/statistics/white-paper-mlit-index.html (accessed on 28 October 2024).
3. Lois, D.; Wang, Y.; Boggio-Marzet, A.; Monzon, A. Multivariate analysis of fuel consumption related to eco-driving: Interaction of driving patterns and external factors. *Transp. Res. Part Transp. Environ.* **2019**, *72*, 232–242. [CrossRef]
4. Sivak, M.; Schoettle, B. Eco-driving: Strategic, tactical, and operational decisions of the driver that influence vehicle fuel economy. *Transp. Policy* **2012**, *22*, 96–99. [CrossRef]
5. Barkenbus, J.N. Eco-driving: An overlooked climate change initiative. *Energy Policy* **2010**, *38*, 762–769. [CrossRef]
6. Barla, P.; Gilbert-Gonthier, M.; Lopez Castro, M.A.; Miranda-Moreno, L. Eco-driving training and fuel consumption: Impact, heterogeneity and sustainability. *Energy Econ.* **2017**, *62*, 187–194. [CrossRef]
7. Ministry of Economy, Trade and Industry, Japan. J-Credit Scheme. 2013. Available online: https://japancredit.go.jp/english/ (accessed on 28 October 2024).
8. Aioi Nissay Dowa Insurance Co. , Ltd. Telematics Auto Insurance: Insured Vehicles Top the 1 Million Mark! 2021. Available online: https://dps.aioinissaydowa.co.jp/iportal/CatalogViewInterfaceStartUpAction.do?method=startUp&mode=PAGE&catalogId=2469390000&pageGroupId=1&volumeID=AIO18001 (accessed on 28 October 2024).
9. Fafoutellis, P.; Mantouka, E.G.; Vlahogianni, E.I. Eco-Driving and Its Impacts on Fuel Efficiency: An Overview of Technologies and Data-Driven Methods. *Sustainability* **2020**, *13*, 226. [CrossRef]
10. Young, R.; Fallon, S.; Jacob, P.; O'Dwyer, D. Vehicle Telematics and Its Role as a Key Enabler in the Development of Smart Cities. *IEEE Sens. J.* **2020**, *20*, 11713–11724. [CrossRef]
11. Ghaffarpasand, O.; Burke, M.; Osei, L.K.; Ursell, H.; Chapman, S.; Pope, F.D. Vehicle Telematics for Safer, Cleaner and More Sustainable Urban Transport: A review. *Sustainability* **2022**, *14*, 16386. [CrossRef]
12. Ghaffarpasand, O.; Pope, F.D. Telematics data for geospatial and temporal mapping of urban mobility: Fuel consumption, and air pollutant and climate-forcing emissions of passenger cars. *Sci. Total Environ.* **2023**, *894*, 164940. [CrossRef] [PubMed]
13. Singh, M.; Dubey, R. Deep Learning Model Based CO_2 Emissions Prediction Using Vehicle Telematics Sensors Data. *IEEE Trans. Intell. Veh.* **2023**, *8*, 768–777. [CrossRef]
14. Maroju, R.; Nishimura, S.; Wang, Z.; Matsuhashi, R. Estimating Vehicular Fuel Consumption and CO_2 Emissions by Machine Learning Using Only Speed and Acceleration. *J. Jpn. Soc. Energy Resour.* **2023**, *44*, 30–38.
15. Ghahramani, M.; Pilla, F. Analysis of Carbon Dioxide Emissions From Road Transport Using Taxi Trips. *IEEE Access* **2021**, *9*, 98573–98580. [CrossRef]
16. Ping, P.; Qin, W.; Xu, Y.; Miyajima, C.; Takeda, K. Impact of Driver Behavior on Fuel Consumption: Classification, Evaluation and Prediction Using Machine Learning. *IEEE Access* **2019**, *7*, 78515–78532. [CrossRef]

17. Rios-Torres, J.; Liu, J.; Khattak, A. Fuel consumption for various driving styles in conventional and hybrid electric vehicles: Integrating driving cycle predictions with fuel consumption optimization. *Int. J. Sustain. Transp.* **2019**, *13*, 123–137. [CrossRef]
18. Abediasl, H.; Ansari, A.; Hosseini, V.; Koch, C.R.; Shahbakhti, M. Real-time vehicular fuel consumption estimation using machine learning and on-board diagnostics data. *Proc. Inst. Mech. Eng. Part J. Automob. Eng.* **2024**, *238*, 3779–3793. [CrossRef]
19. Schoen, A.; Byerly, A.; Hendrix, B.; Bagwe, R.M.; dos Santos, E.C.; Ben Miled, Z. A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles. *IEEE Trans. Veh. Technol.* **2019**, *68*, 6343–6351. [CrossRef]
20. Sai Manvitha, M.; Vani Pujitha, M.; Hari Prasad, N.; Yashitha Anju, B. A Predictive Analysis on CO_2 Emissions in Automobiles using Machine Learning Techniques. In Proceedings of the 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 5–7 January 2023; pp. 394–401.
21. Hattori, M.; Shimizu, O.; Nagai, S.; Fujimoto, H.; Sato, K.; Takeda, Y.; Nagashio, T. Quadrant Dynamic Programming for Optimizing Velocity of Ecological Adaptive Cruise Control. *IEEE/ASME Trans. Mechatronics* **2022**, *27*, 1533–1544. [CrossRef]
22. Cournapeau, D. scikit-learn 1.0.2. 2021. Available online: https://scikit-learn.org (accessed on 28 October 2024).
23. Python Software Foundation. Python 3.7.13. 2022. Available online: https://www.python.org (accessed on 28 October 2024).
24. Wang, Z.; Mae, M.; Nishimura, S.; Matsuhashi, R. Vehicular Fuel Consumption and CO_2 Emission Estimation Model Integrating Novel Driving Behavior Data Using Machine Learning. *Energies* **2024**, *17*, 1410. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.