Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230842

Prediction of carbon dioxide emissions based on machine learning algorithms

Zhihuang Chen

The department of Computer Science, California State University, Fullerton, Fullerton, CA 92831-3599, the United States

zhchen@csu.fullerton.edu

Abstract. The escalating emergence of environmental issues, including the greenhouse effect and coral bleaching, has raised global awareness of the significance of sustainable development and preservation of the earth's resources. Although reducing emissions is essential to mitigate the adverse effects of greenhouse gases, it remains challenging to detect without specialized equipment. This constraint is particularly burdensome for small organizations and individual groups due to the high associated costs. Therefore, this study proposes using machine learning algorithms and common vehicle attributes to predict greenhouse gas emissions accurately. Specifically, the research employs the random forest algorithm, incorporating vehicle power parameters to predict carbon dioxide emissions. The study employs Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared metrics to analyze the model's effectiveness, accuracy, and feasibility in predicting greenhouse gas emissions. This approach will enable small groups to participate in environmental protection efforts, democratizing the process for all who desire to safeguard the environment.

Keywords: machine learning, random forest, prediction of carbon dioxide emissions.

1. Introduction

With climate change caused by environmental problems worldwide, individuals are increasingly aware of the need for environmental protection. A 2020 survey from Ipsos found that 71% of people globally believe the world is facing a climate emergency, up from 58% in 2019 [1]. Among them, carbon dioxide is one of the main gases causing environmental problems. From the Issues and Local Programs page on the Washington State Department of Ecology's official website, Carbon dioxide emitted by vehicles is the most common anthropogenic greenhouse gas, according to an article by the Washington State Department of Ecology advocating for environmental protection [1]. In other words, humans can effectively reduce corresponding environmental problems by reducing the carbon emissions of vehicles. Today, relevant departments and automakers in many countries have taken up the challenge of reducing vehicle carbon emissions by analyzing data from sensors and other sources combined with machine learning to predict and optimize a vehicle's carbon emissions.

Machine learning, a subfield of artificial intelligence that uses algorithms and statistical models that allow computer systems to learn and improve from experience, has already significantly reduced carbon emissions from vehicles. For example, machine learning can optimize the vehicle's engine parameters to reduce carbon emissions by analyzing the patterns and correlations among factors such as vehicle

DOI: 10.54254/2755-2721/15/20230842

performance, weather, traffic conditions, and driving habits. This data can help vehicle manufacturers develop more economical and environmentally friendly vehicles. Machine learning has undeniable potential to help reduce vehicle carbon emissions. By analyzing data from sensors and other sources, machine learning algorithms can optimize vehicle performance, reduce fuel consumption, and minimize carbon emissions. As the automotive industry continues to adopt machine learning, it's not hard to imagine a significant reduction in carbon emissions from cars in the coming years. At the same time, many companies have also launched projects and services that optimize vehicles through machine learning algorithms, such as Tesla [2], Ford [3], Toyota [4], BMW [5], and Volkswagen [6]. Despite the benefits of this method, a limitation exists in that the prediction results heavily depend on the accuracy and availability of sensor data. Obtaining and predicting carbon emission data is expensive for some small businesses, organizations, and the public without professional equipment. Protecting the earth's ecological environment is crucial to the sustainable development of human beings. It requires the joint efforts of all human beings, and the cost and hardware requirements make it more difficult for most people to participate in protecting the earth's ecological environment. Therefore, developing low-cost and accurate methods for predicting vehicle carbon emissions without professional equipment is an important research topic to promote broader participation in environmental protection.

This study aims to study the feasibility of predicting vehicle carbon emissions without relying on sensors based on vehicle data recorded by government departments and authoritative organizations combined with machine learning algorithms. In order to prevent unnecessary misleading to the users of the research results and maintain the objectivity of the research data, this research will not include the brand, model, vehicle manufacturing process, and technology of the vehicle. However, the selection of the characteristics of the project will take the dynamics of the vehicle and its basic characteristics as the main parameters of the study. Considering that the actual carbon emissions of vehicles may vary depending on the influence of uncertain factors such as vehicle parameters, driving environment, and driving habits, this research's machine learning algorithm model will use ensemble learning. The main advantages of ensemble learning over traditional single-model approaches are its flexibility and scalability. It can help reduce the impact of uncertainty and variability in data in situations where there is a high degree of uncertainty or variability in the data. In addition, it will make the results of this research more informative.

2. Method

2.1. Dataset preparation

In this study, the scikit-learn (sklearn) toolkit was utilized to process, analyze and evaluate the data. In terms of data collection, in order to ensure the authenticity and accuracy of the data. The data used in this study includes vehicle parameters and carbon emission data released by government departments and authoritative institutions. The sample data size is 12, 988, including common vehicles from 2019 to 2023 [7]. Among them, 11,589 were for conventional power vehicles (i.e. gasoline and diesel), 712 were for electric vehicles, 46 were for hydrogen-powered vehicles, 424 were for gasoline-electric hybrid vehicles, 424 were for biofuels, and 217 were for electric hybrid vehicles (i.e. fuel ethanol and electricity). Due to the different fuels the vehicles use, carbon emissions, and various attributes vary widely and go unrecorded. For example, the carbon emissions of pure electric vehicles are always zero, and the engine parameters of biofuel vehicles are not well documented. Considering that these factors may have adverse effects on the results, and this study aims to accurately predict the feasibility of vehicle carbon emissions by analyzing data without professional sensors, this study will only use conventionally powered vehicles as Training data, including gasoline and diesel vehicles. Therefore, the actual usage data is 11, 589.

Furthermore, in order to identify meaningful parameters in the data. Linear regression was used to analyze the relationship between vehicle parameters and CO2 emissions. In the obtained results, it is found that there is a certain linear relationship between the carbon dioxide emission of the vehicle and the power and fuel consumption of the vehicle. Therefore, in the feature selection of the model, these

DOI: 10.54254/2755-2721/15/20230842

data with a linear relationship are used as the main imported data. For some missing data, this study uses a module in the scikit-learn library to perform mean imputation on missing data. Consider that the ordering of the data in the original data is random. Therefore, using the mean value of the data to replace missing values can keep the overall distribution characteristics of the data unchanged, thereby minimizing the error caused by missing data. For data type conversion, based on the algorithm features used in this research, the data type is uniformly converted to real numbers (i.e. float).

2.2. Machine learning models

This study aims to predict vehicle carbon emissions through vehicle parameters. Given that the cited data contains missing and uncertain values, the algorithm of choice for this study is the random forest algorithm within the supervised learning machine learning framework [8-10]. Random forest is an ensemble learning method based on decision trees and is widely used in machine learning due to its ability to enhance model accuracy and stability by combining multiple decision trees. Also, random forests are less susceptible to noisy data. It goes through a decision forest consisting of multiple decision trees, each consisting of randomly selected features and samples. At the same time, the random forest algorithm can deal with missing data and effectively deal with missing values in the data set. The working principle of the random forest is that its training set is generated by bootstrap sampling, and features are randomly selected in the new training set to build a decision tree. Using each decision tree to predict new sample data after building a certain number of decision trees. The results from each tree are then aggregated using voting or averaging methods to arrive at a final prediction. Random forest reduces the variance of the decision tree by randomly selecting features and samples, thereby improving the generalization ability of the overall model. Therefore, the advantage of using the random forest algorithm in this study is that it can effectively reduce the negative impact of a series of uncertain factors, such as different technologies of automobile manufacturers, thereby improving the accuracy of the research results.

3. Result and discussion

Through random forest algorithm model analysis, the mean square error obtained in this study is 40.4801, the root mean square error is 6.3624, and the R-squared is 0.9961 (Figure 1). The possible values of the target variable (Comb CO2) ranged from 151 to 979, with a mean of 407.2 (Figure 2). A Mean Square Error (MSE) of 40.48 indicates that the model's predictions are off by about 6.36 units (average square root of 40.48). This value appears relatively small compared to the range of possible values for the target variable.

Second, the Root Mean Square Error (RMSE) of 6.36 is also relatively small compared to the mean and range of possible values of the target variable, which indicates the model's performance. Also, an R-squared value of 0.996 indicates that the model can explain most of the variance in the target variable. In other words, the model's predictions are highly correlated with the actual value of the target variable. It can be seen from Figure 2 that the direct relationship between the predicted value and the actual value is linear, and the image is close to a straight line. This indicates that the model's predictions are similar or the same as the actual values. Furthermore, most of the deviations in the predicted values of the model are concentrated around zero, as illustrated in Figure 3. This finding indicates that the model's predicted results closely match the actual values, underscoring its suitability for reference in this study's data analysis.

Overall, a MSE of 40.48 is deemed acceptable for this study, indicating that the model's predictions are usually accurate. Secondly, the RMSE of 6.36 is also within the acceptable range, the deviation is small, and the model has an excellent R-squared value (0.996). Based on these metrics and the context, the random forest model performs remarkably well on this dataset. Therefore, predicting the vehicle's carbon emissions through the random forest algorithm model is feasible.

DOI: 10.54254/2755-2721/15/20230842

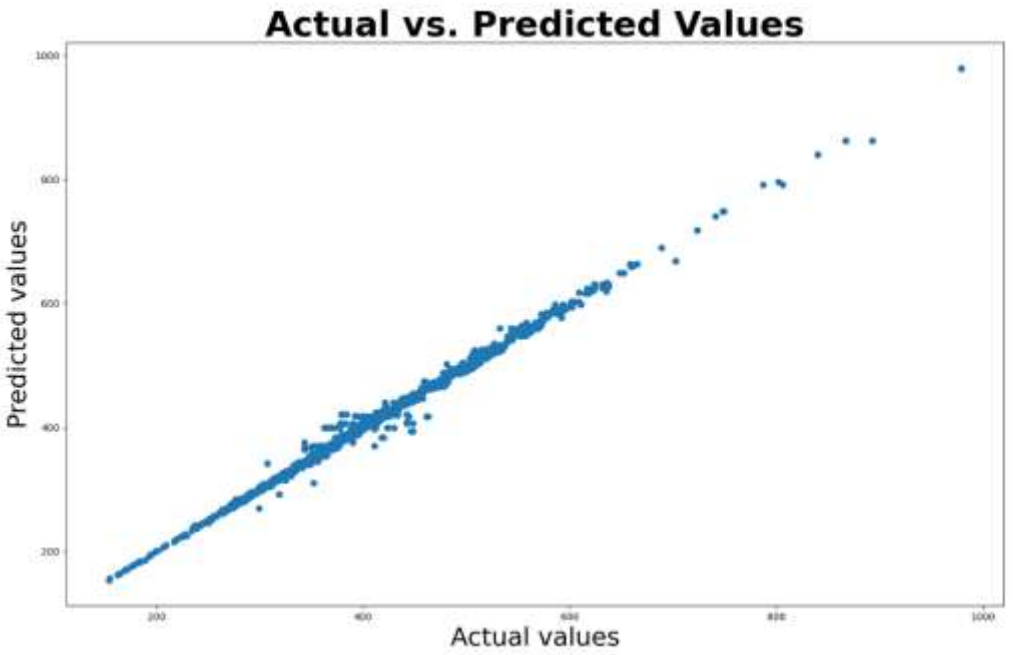

Figure 1. The relationship between actual and predicted CO2 emissions.

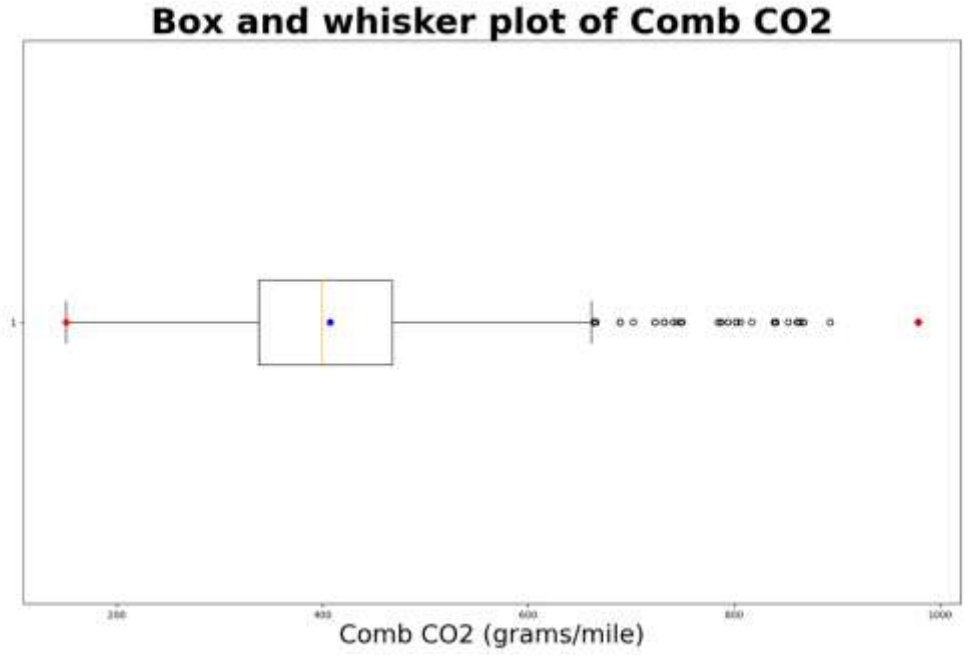

Figure 2. The statistics for the sample data.

DOI: 10.54254/2755-2721/15/20230842

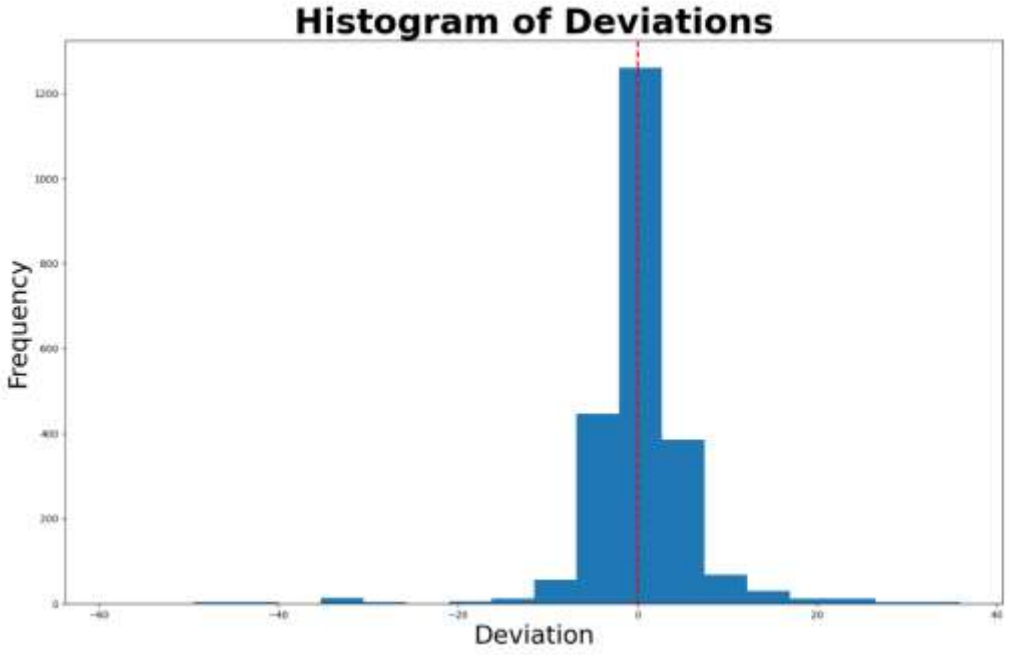

Figure 3. The deviation value of the predicted data.

4. Conclusion

In this study, the purpose of the research is to test the hypothesis that there is a way to help all those who want to participate in environmental protection to reduce the limitations and costs. In the experiment, carbon dioxide, one of the common greenhouse gases, was used as the predicted target for research and analysis. The results of this experiment show that it is feasible to predict carbon dioxide emissions by combining common vehicle attributes with the random forest algorithm. Furthermore, this study's MSE, RMSE, and R average prove that the sample data fit the algorithm model well. It also means that the method has the potential to predict other greenhouse gases. For subsequent improvements, it will be considered to apply the algorithm to the prediction of other greenhouse gases and add some factors that may affect the prediction results, such as the year of the vehicle, road conditions, and the type of vehicle fuel.

References

[1] Reducing air pollution from cars 2023 Reducing car pollution - Washington State Department of Ecology Retrieved April 5 2023 from https://ecology.wa.gov/Issues-and-local-projects/Education-training/What-you-can-do/Reducing-car-pollution

[2] AI & Robotics 2023 Tesla. Retrieved April 5, 2023, from https://www.tesla.com/AI

[3] Ford Establishes Latitude AI to Develop Future Automated Driving Technology 2023 Ford Media Center. Retrieved April 5, 2023, from https://media.ford.com/content/fordmedia/fna/us/en/news/2023/03/02/ford-establishes-latitude-ai-to-develop-future-automated-driving.html

[4] Toyota automated driving. 2023 Retrieved April 5, 2023, from https://amrd.toyota.com/app/uploads/2022/02/ATwhitepaper.pdf

[5] BMWK - Federal Ministry for Economics Affairs and Climate Action. 2023 AI-based solution for optimizing the energy efficiency and consumption of electric vehicles BMWK Retrieved

DOI: 10.54254/2755-2721/15/20230842

April 5, 2023, from https://www.bmwk.de/Redaktion/EN/Artikel/Digital-World/GAIA-X-Use-Cases/77-gaia-x-decentralized-in-vehicle-mlaas-to-ev-energy-efficiency/use-case.html
[6] Volkswagen autonomous driving in Hamburg 2022 Volkswagen Group Retrieved April 5, 2023, from https://www.volkswagenag.com/en/news/stories/2019/04/laser-radar-ultrasound-autonomous-driving-in-hamburg.html
[7] Energy 2023 Fuel Economy Data https://www.fueleconomy.gov/feg/download.shtml
[8] Biau G Scornet E 2016 A random forest guided tour Test 25: 197-227
[9] Rigatti S J 2017 Random forest Journal of Insurance Medicine 47(1): 31-39
[10] Oshiro T M Perez P S Baranauskas J A 2012 How many trees in a random forest? Machine Learning and Data Mining in Pattern Recognition: 8th International Conference MLDM 2012 Berlin Germany July 13-20 Proceedings 8 Springer Berlin Heidelberg 154-168