Machine Learning-Based Prediction of Carbon Dioxide Emissions from Automobiles and Influencing Factors

Yuxin Liu *

Department of Math, Xi'an University of Finance and Economics, Xi'an, China

* Corresponding Author Email: liuyuxin125@xaufe.edu.cn

Abstract. In recent years, the number of automobiles has been increasing globally, further leading to an increase in carbon dioxide emissions year after year. Carbon dioxide emissions from automobiles have become an important factor in global climate change and have attracted global attention. Firstly, the data were analyzed for missing values all duplicates were removed, and the main discrete random variables were converted into continuous random variables; then, the correlation coefficients of Pearson correlation coefficients were used to analyze the correlation between each automobile characteristic, and heat maps were drawn to remove the influencing factors with weak correlation; finally, the prediction models were established based on multiple linear regression and the Random Forest method, respectively. The results show that in both models, Fuel Consumption Comb is the most important factor influencing the growth of automobile carbon dioxide emissions; the random forest model is better than the multiple linear regression model and can effectively predict automobile carbon dioxide emissions.

Keywords: Machine earning, carbon dioxide emissions, multiple regression, random forest.

1. Introduction

Global warming, which is a result of the greenhouse effect, has grown to be an international issue as a result of societal advancement and economic expansion [1]. Carbon dioxide gas has become the main component affecting the greenhouse effect, accounting for about two-thirds of greenhouse gases [2]. The number of automobiles around the world is increasing, which further leads to an increase in carbon dioxide gas emissions year after year.

The automobile has become an indispensable part of people's lives, but it is also accompanied by huge energy consumption and environmental pollution. Carbon dioxide emissions from automobile exhaust have become the main source of global carbon dioxide emissions. Carbon dioxide emissions from automobiles will directly affect the global climate, causing problems such as rising global temperatures, rising sea levels, and climate anomalies. In order to prevent global warming and climate change from causing irreversible impacts on ecosystems, biodiversity, and human societies, governments have put in place a variety of policy initiatives to reduce carbon dioxide emissions from vehicles.

Hence, the research topic of this paper focuses on solving the problem of predicting and influencing factors of vehicle carbon dioxide emissions.

2. Literature Review

In order to explore the issue of carbon dioxide emissions, researchers have developed several analytical models over the years.

Pan Siyu and Zhang Meiling used a neural network algorithm to establish a BP neural network model based on the impact of historical carbon dioxide emissions to forecast the trend of carbon dioxide emission time series in Gansu Province [3]. Yan Shiyang, Liu Huilin, Mo Zhaoyu, et al. Constructed a Long-range Energy Alternatives Planning System model to predict carbon dioxide emissions in the power sector [4]. Zuo Qiting, Zhao Chenguang, Ma Junxia, et al. proposed the carbon dioxide emission equivalent analysis of water resource behavior from the four dimensions of water resource development, allocation, utilization, and protection and constructed a function table for

carbon dioxide emission equivalent analysis [5]. Jiang Jianan used VAR modeling to explore the dynamic impacts of carbon dioxide emission reduction in the cement industry in Hebei Province by taking economic growth, energy productivity, energy structure, carbon productivity, and other influencing factors as variables [6].

To forecast carbon dioxide emissions, they all employ various techniques. The carbon dioxide emissions in this report will be predicted using machine learning.

3. Methodology

This report focuses on the use of machine learning. Machine learning, based on data, builds models using the principles of different algorithms and then trains them repeatedly. By conducting an automated analysis of the data, machine learning models can identify implicit laws between input variables and target attributes, and then predict target properties based on these rules. Machine learning can be divided into supervised learning, non-surveillance learning, and semi-supported learning depending on whether the training process has a given target value. Supervised Learning has a target value in the learning process, and the output data of model training is often used to solve regression and classification problems [7].

The experimental steps in this report can be roughly divided into four parts. It contains data description, preprocessing data, correlation analysis visualization modeling, and forecasting.

3.1. Data Description

The data has been taken and compiled from the official Canada Government link (https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6). This contains data over a period of 7 years. There are a total of 7385 rows and 12 columns. There are a few abbreviations that have been used to describe the features. The specific influencing factors are shown in Table 1.

Table 1. Vehicle Characteristics Information

Variable name	Feature type	Characteristic Implications
Make	Discrete	Company of the vehicle
Model	Discrete	Car model
Vehicle Class	Discrete	Class of vehicle depending on their utility, capacity and weight
Engine Size	Continuous	Size of engine used in liters
Cylinders	Discrete	Number of cylinders
Transmission	Discrete	Transmission type with number of gears
Fuel type	Discrete	Type of Fuel used
Fuel Consumption City	Continuous	Fuel consumption in city roads (L/100 km)
Fuel Consumption Hwy	Continuous	Fuel consumption in Hwy roads (L/100 km)
Fuel Consumption Comb	Continuous	The combined fuel consumption (55% city, 45% highway) is shown in L/100 km
Fuel Consumption Comb mpg	Continuous	The combined fuel consumption in both city and highway is shown in mile per gallon(mpg)

3.2. Preprocessing Data

First, simple processing of the data shows that there are no missing values, but there are 1103 lines duplicating, and in order to reduce the impact of duplication, this experiment decides to delete the duplicate values. Secondly, the experimental data contains four floating-point data points, three integer data points, and five objective data points. Then, the categorical data is encoded into numerical data, and the types of fuels from X, Z, E, D, and N are replaced with 1, 2, 3, 4, and 5.

3.3. Correlation Analysis and Visualization

Pearson's correlation analysis is a common descriptive analysis method designed to examine correlations between characteristics. The Pearson correlation coefficient is a value between 1 and -1, indicating a linear relationship between two variables. If one variable is increased, the other variable also increases, indicating that the two variables are positively correlated, with a correlation coefficient greater than 0. If one variable increases, the other is decreased, which indicates that there are negative correlations, that are less than 0 [8]. The formula for calculating the Pearson coefficient is:

$$\rho_{x,y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{cov(X,Y)E[(X-\mu x)(Y-\mu y)]}{\sigma_x \sigma_y} \quad (1)$$

When the absolute value of the associated coefficient is between 0.8 and 1.0, the indication of the two characteristics is highly correlated; when the absolutes of the related coefficient are between 0.6 and 0.8, the indicator is strongly correlative; when an absolute of the relevant coefficients is between 0.4 and 0.6, it is indicative of the medium degree of correlation of both characteristics; when a relative coefficient is between 0.2 and 0.4, it indicates the weak correlations between the two features; and when the absolutes of the corresponding coefficient are between 0.0 and 0.2, the indicators are very weak or irrelevant.

Calculate the correlation coefficient between the characteristics of the vehicle and draw a heat map, as shown in Fig. 1.

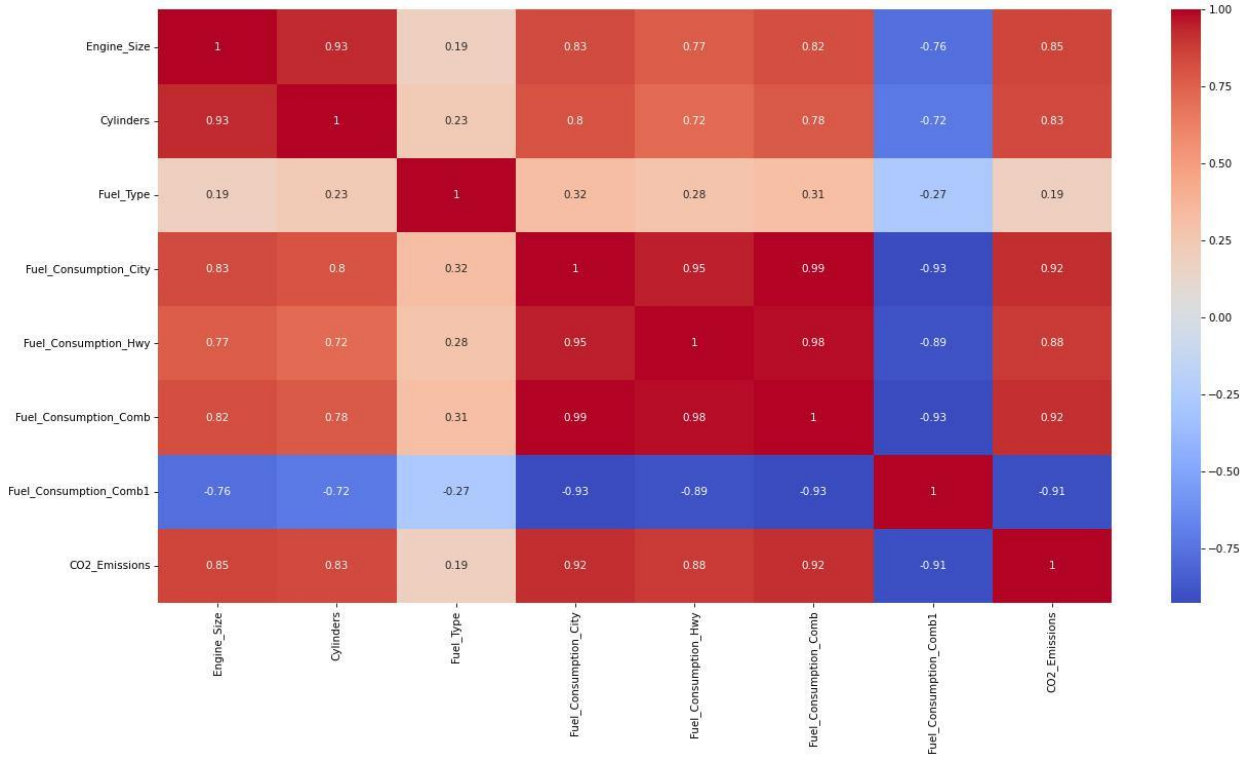

Figure 1. Correlation between Vehicle Characteristics

The heat map shows a very weak correlation between fuel type and CO2 emissions, from which it is known that fuel type is not the main factor affecting the vehicle's CO2 emissions; the correlated coefficient between Fuel Consumption Comb1 and CO2 emissions is less than 0, which is strongly negative; and Engine Size, Cylinders, Fuel Consumption City, Fuel Consumption Hwy, and Fuel Consumption Comb are strongly correlated with CO2, so that Engine Size, Cylinders, Fuel Consumption City, Fuel Consumption Hwy, and Fuel Consumption Comb may be the major factor influencing car CO2 emissions.

4. Modeling and forecasting

4.1. Multiple Linear Regression Model

If you follow the "checklist" your paper will conform to the requirements of the publisher and facilitate a problem-free publication process. Multiple linear regression is an important method in multiple statistical analyses. In many practical problem studies, because variable changes are often influenced by several important factors, it is necessary to use two or more influence factors as self-variables to explain variable-related changes, that is, multiple regression [9]. The linear regression model of random variable y and general variable $x_1, x_2, ..., x_p$ is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2}$$

In this report, the unknown parameter $\beta_0, \beta_1, ..., \beta_p$ in the multiple linear regression equation is estimated using the least squares. The five influence factors of Engine Size, Cylinders, Fuel Consumption City, Fuel Consumption Hwy, and Fuel Consumption Comb as self-variables, and CO2 emissions as a relative variable. Build a multiple linear regression model using Python, as shown in Fig. 2

Dep. Variable:	CO2_Emissions	**R-squared:**	0.879
Model:	OLS	**Adj. R-squared:**	0.879
Method:	Least Squares	**F-statistic:**	6403.
Date:	Thu, 16 Nov 2023	**Prob (F-statistic):**	0.00
Time:	08:59:31	**Log-Likelihood:**	-19510.
No. Observations:	4397	**AIC:**	3.903e+04
Df Residuals:	4391	**BIC:**	3.907e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.0815	1.537	33.877	0.000	49.067	55.096
Engine_Size	6.1038	0.671	9.099	0.000	4.789	7.419
Cylinders	6.2136	0.466	13.330	0.000	5.300	7.127
Fuel_Consumption_City	-2.3117	4.183	-0.553	0.581	-10.513	5.890
Fuel_Consumption_Comb	17.3690	7.594	2.287	0.022	2.481	32.257
Fuel_Consumption_Hwy	-1.9119	3.458	-0.553	0.580	-8.691	4.867

Omnibus:	1235.397	**Durbin-Watson:**	1.953
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4301.028
Skew:	-1.386	**Prob(JB):**	0.00
Kurtosis:	6.973	**Cond. No.**	627.

Figure 2. OLS Regression Results

As shown in Figure 2, the multiple linear regression model of CO2 emissions is as follows:

$$\begin{aligned} CO2\ emissions &= 52.08 + 6.10 \times Engine\ Size + 6.21 \times Cylinders \\ &- 2.31 \times Fuel\ Consumption\ City + 17.37 \times Fuel\ Consumption\ Comb \\ &- 1.91 \times Fuel\ Consumption\ Hwy \end{aligned}$$

First, Prob(F-statistic) = 0.00, stating that the model is reliable. Second, the determination factor of the analytical model is R-squared = 0.879, indicating a better degree of adaptation of the regression model. Except for variables Fuel Consumption City and Fuel Consumption Hwy, the P values of the T-test p>|t| are all smaller than 0.05, indicating that two self-variables Fuel Consumption City and

Fuel Consumption Hwy are not noticeable and the model needs further optimization. Third, Durbin-Watson = 1.953, close to 2, indicating the residues of regression models have basically no self-relevance. Fourth, Cond. No. is used to measure the presence of multicollinearity between the independent variables of the multiple regression model. Cond. No.= 627 indicates that there is more multicollinearity between the independent variables.

Since the established linear regression model contains two insignificant characteristics, it is possible to re-establish the multiple linear regression model after removing the insignificant characteristics. The new multiple linear regression model of CO2 emissions is as follows:

$$CO2\ emissions = 52.05 + 6.10 \times Engine\ Size + 6.21 \times Cylinders + 13.15 \times Fuel\ Consumption\ Comb$$

$R-squared$ remains equal to 0.879; the P values $p > |t|$ for all the T-tests are less than 0.05, indicating that all self-variables at this time are significant; $Cond.\ No. = 65.2$, which is less than 100, indicates that the multicollinearity between the independent variables is small.

4.2. Random Forest Regression Model

Random Forest is a flexible and easy-to-use machine learning algorithm that contains multiple decision tree combination classifier algorithms. Random forest has obvious advantages in dealing with multidimensional data and is one of the best classification algorithms available. It uses the bootstrap method to extract k samples from the original training sample set N and then builds the corresponding decision-tree model for each of the samples. Finally, a vote was held on the results of the k samples obtained, and the final classification result was selected on the basis of the principle of minority obedience to the majority [10]. The classification decision function is:

$$H(x) = \arg\max y \sum_{i=1}^{k} I[h_i(x) = Y] \quad (3)$$

Calculate the importance of characteristics for the establishment of the random forest on the basis of data; see Fig. 3

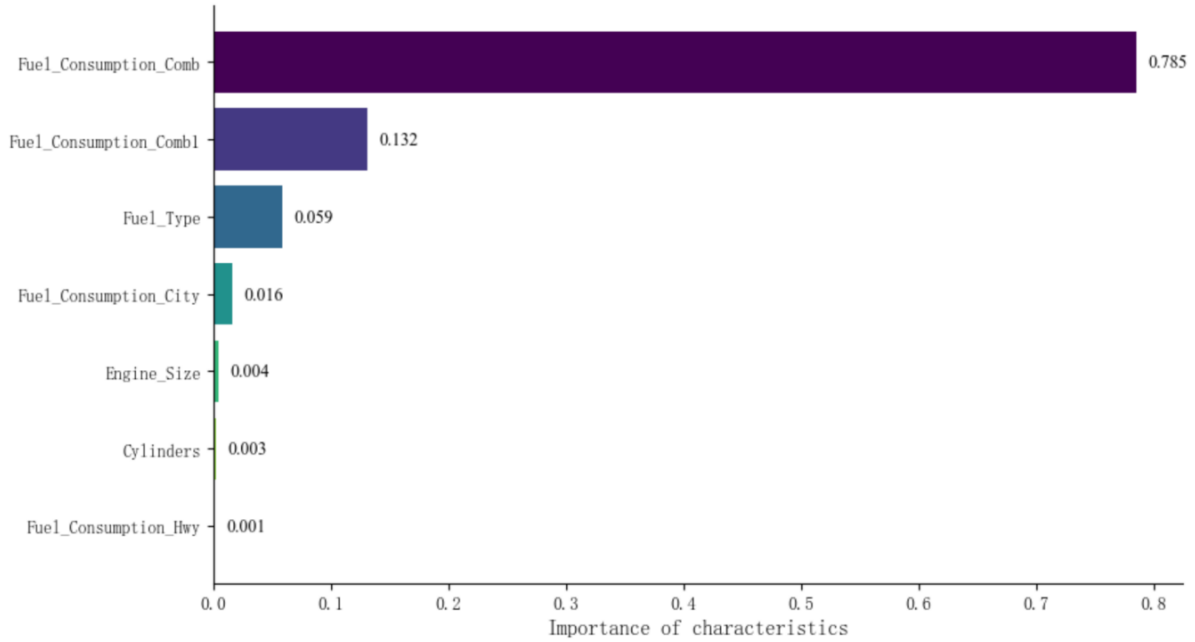

Figure 3. Visualization of the Importance of Random Forest Characteristics

By analyzing the data, the characteristic value of the Fuel Consumption Comb is 0.785, the characteristic value of Fuel Consumption Comb1 is 0.132, and it can be concluded that the main factor affecting the car's CO2 emissions is the fuel consumption comb. Fuel Consumption Comb1

also has some influence factors, but there is a negative correlation between fuel Consumption Comb1 and CO2 emissions.

A random forest regression model was established, with 70% divided into training sets and 30% into test sets. The accuracy of training sets at this time was 99.54%, and test sets were 97.93%. For an intuitive view of the predicted and true values of the sample, the forecast and real values were visualized using the slide chart. As shown in Fig. 4.

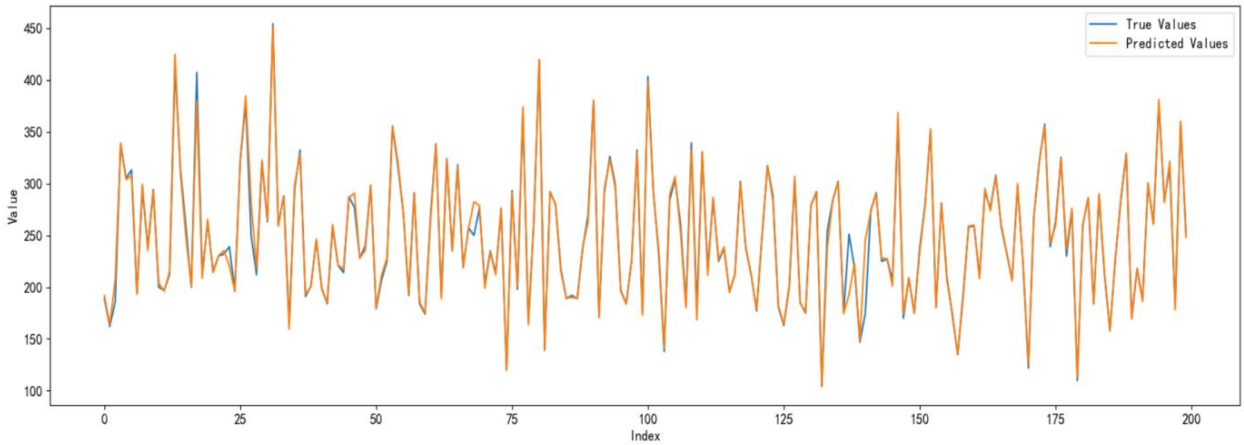

Figure 4. Comparison diagram of predicted and true values

Figure 4 shows that most samples can be predicted accurately, and only a small number of samples have small errors in the predictive and true values, indicating that the model predicts better.

The authors gratefully acknowledge the financial support from xxx funds.

5. Conclusion

In this report, multiple linear regression model and random forest model are used to predict the carbon dioxide emission of automobiles and analyze the main factors affecting the carbon dioxide emission of automobiles. The experiments show that in both models, Fuel Consumption Comb is the most dominant influencing factor, and Fuel Consumption Comb1 is also one of the influencing factors, but with a negative correlation. Meanwhile, the random forest model performs better than the multiple linear regression model.

References

[1] Xue Yunfei. Research on prediction of automobile carbon dioxide emissions based on machine learning. Automation & Information Engineering, 2023, 44 (1): 22 - 26; 45.

[2] Chen Ruilang, Qiao Yongping. Carbon dioxide gas from automobile exhaust emissions. Science and Technology Information, 2012, (26): 99, 101.

[3] Pan Siyu, Zhang Meiling. Research on the prediction of carbon dioxide emission and influencing factors in Gansu Province based on BP neural network. Environmental Engineering, 2023, 41 (7): 61 - 68, 85.

[4] Yan Shiyang, LIU Huilin, MO Zhaoyu et al. Development status and carbon dioxide emission forecast of power industry in Guangxi. Popular Science and Technology, 2023, 25 (5): 23 - 27.

[5] Zuo Qiting, Zhao Chenguang, Ma Junxia et al. Carbon dioxide equivalent analysis of water resources behavior and its application. South-to-North Water Diversion and Water Resources Science and Technology (in Chinese and English), 2023, 21 (1): 1 - 12.

[6] Jiang Jianan. Research on the influencing factors of carbon dioxide emission of cement industry in Hebei Province based on VAR model. Modern Commerce Industry, 2023, 44 (9): 267 - 268.

[7] Song Weixue. Research on CO2 emission prediction model of construction industry based on machine learning. Xi'an University of Architecture and Technology, 2020.

[8] Wang Jinwei, SAN Yuhan, SAN Baohai. Analysis of short track speed skating competition data based on Pearson's correlation coefficient. Ice and Snow Sports, 2023, 45 (4): 9 - 12, 17.

[9] Guo Juan. Multiple linear regression automobile fuel consumption prediction model based on the existence of interaction terms. Guangxi quality supervision guide, 2019, (11): 138 - 140.

[10] Xiao Jinjuan, Pang Jinxiang, Chen Wenzhuo. Principal component identification and classification of ancient glass based on random forest model. Science and Technology Innovation, 2023, (14): 37 - 40.