IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 29, 2019, accepted April 25, 2019, date of publication May 1, 2019, date of current version May 28, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2914378*

# Multilayer Perceptron Method to Estimate Real-World Fuel Consumption Rate of Light Duty Vehicles

**YAWEN LI[1], GUANGCAN TANG[2], JIAMENG DU[3], NAN ZHOU[4], YUE ZHAO[5], AND TIAN WU[6,7,8]**

[1]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong
[3]Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 96801, USA
[4]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[5]International School, Beijing University of Posts and Telecommunications, Beijing 100876, China
[6]NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[7]School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[8]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Tian Wu (wutian@amss.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71804181, in part by the National Key Research and Development Program of China under 2018YFC0807205, in part by the National Center for Mathematics and Interdisciplinary Sciences, CAS, and in part by the Fundamental Research Funds for the Central Universities.

**ABSTRACT** The actual driving condition and fuel consumption rate gaps between lab and real-world are becoming larger. In this paper, we demonstrate an approach to determine the most important factors that may influence the prediction of real-world fuel consumption rate of light-duty vehicles. A multilayer perceptron (MLP) method is developed for the prediction of fuel consumption since it provides accurate classification results despite the complicated properties of different types of inputs. The model considers the parameters of external environmental factors, the manipulation of vehicle companies, and the drivers' driving habits. Based on the BearOil database in China, 2,424,379 samples are used to optimize our model. We indicate that differences exist between real-world fuel consumption and standard fuel consumption under simulation conditions. This study enables the government and policy-makers to use big data and intelligent systems for energy policy assessment and better governance.

**INDEX TERMS** Artificial intelligence, big data, multilayer perceptron, fuel consumption rate, light-duty vehicles.

## I. INTRODUCTION

In order to reduce pollution and protect air quality, the promotion of green energy is currently gaining increasing attention in China [19]. In recent years, China has issued a series of policies to actively encourage the production and consumption of energy-saving and new energy vehicles. The Ministry of Industry and Information Technology (MIIT), in conjunction with the National Development and Reform Commission, the Ministry of Commerce, the General Administration of Customs and the General Administration of Quality Supervision and Inspection, have drafted the Notice on Strengthening the Management of Average Fuel Consumption in Passenger Vehicle Enterprises in 2014. Furthermore, to achieve the overall goal of the average fuel consumption for passenger vehicles, which will be 5 L/100 km in 2020, MIIT has released the revised version of the Limits of Fuel Consumption for Passenger Vehicles and Evaluation Method and Index of Fuel Consumption of Passenger Vehicles, which states that China will formally implement the fourth phase of fuel consumption standards from January 1, 2016. From 2016 to 2020, the set goal of average fuel consumption for passenger vehicles is 6.7 liters, 6.4 liters, 6 liters, 5.5 liters and 5 liters, respectively, for each year. The standard fully meets the current level of EU standards for automotive oil products, and several indexes even exceed the EU standards.

Although great efforts have been made, the automobile industry in China is facing the tremendous pressure of emission reduction [1]. Generally, the driving cycle fuel consumption data are derived from the website of Automobile Fuel

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

Consumption of China by MIIT, which presents a nearly objective value in the numerical simulation conditions with driving operation templates and reflects the reference samples of real-world fuel consumption data under normal circumstances. It has been argued that the standard itself has many limitations and vulnerabilities. Because there is a huge difference between the driving conditions in the test procedure for the automobiles and the real world, the automobile manufactures can easily generate very "perfect" data to meet the standard.

There are three main reasons for the difference between standard fuel consumption and actual fuel consumption. First, the external environmental factors, including the heterogeneity of temperature and atmospheric pressure in different regions, and the degree of road traffic congestion [17]. As urban road traffic congestion is becoming increasingly serious, car idling may lead to an increase in fuel consumption even if the mileage does not change. Second, another reason for the difference between standard fuel consumption and actual fuel consumption is the manipulation of companies. As automobile companies are familiar with the driving cycle test conditions, they usually provide the vehicles with the best conditions for testing to reduce fuel consumption levels. The third is driving habits, such as air conditioning habits and idle parking. Compared to experienced drivers, new drivers always lack energy-saving abilities during the driving process. Due to the increasing number of vehicle drivers, energy saving is facing more and more severe challenges. As a result, even if a car meets the level of emission standard, it could still be harmful to the environment. Thus, we need to explore the most important factors which may influence the prediction of real-world fuel consumption rate of light duty vehicles.

The rest of the paper is organized as follows: Section II reviews the related literature; Section III presents the methodology and data; Section IV analyzes the results of the multilayer perceptron (MLP) and discusses the policy implications; and finally, the conclusion is presented in Section V.

## II. LITERATURE REVIEW

With the rapid development of artificial intelligence (AI), AI-driven big data processing technologies have been used in the field of business activity prediction [8] [9] [24]. During the development of Plug-in hybrid vehicles (PHEVs), big data and AI models also contributed greatly to estimating the electric grid power demand [16]. A study helped promote the development of public charging infrastructure by estimating the refueling demand mined from big-data [3]. Some scholars proposed a range estimation framework through collecting data from the real world to help electric car drivers calculate the remaining driving range [15]. Furthermore, recent work has discussed the challenges of using big data to solve transportation problems, such as the difficulties in collecting and cleaning the data with rich information [18]. Previous studies also found that intelligent systems can be used in calculating electric vehicle charging demand with the help of historical traffic data and weather data [2]. Scholars aim to determine the important factors that may predict the innovation efficiency of a NEVs firm. They have built several machine learning models to help firms increase innovation efficiency and build intelligent decision support systems [9].

As air-quality protection and energy consumption reduction is gaining attention from emerging economies, an International Vehicle Emissions (IVE) Model was developed to estimate the mobile source emissions in an urban area [4]. Another IVE model was designed to evaluate the emissions of light duty gasoline trucks, heavy duty gasoline vehicles and motorcycles by utilizing a dataset in China [6]. To help the government set fuel economy standards for automobiles and reduce energy consumption, scholars have attempted to use the AI-driven (artificial intelligence) big data processing method to create a better test procedure for automobiles, taking the driving situations and environment of the country into consideration. The method can be directly applied to different types of vehicles worldwide [13] [14].

The government and policy-makers have noticed the important role of big data and intelligent systems for policy assessment and better governance. A previous study built a model to calculate the possibility of an automobile industry in achieving the goal of the European Commission's voluntary agreement under baseline conditions [22]. By collecting data from navigation systems, scholars found it possible to analyze the driving and mobility patterns in Europe and calculate the potential of electric vehicles in replacing fuel vehicles. Thus, policies that encouraged the development of the next generation of green vehicles could be implemented [5].

To build next-generation intelligent transportation systems, data processing and mining techniques need to be improved to make full use of real-time information [10] [23]. In this study, we present an efficient AI-driven big data processing method to fill this gap, which can better reflect real-world driving conditions. As previous studies indicate that city characteristics, road infrastructure and driving behavior are the most important factors that may influence vehicle emissions, we believe that this study may help government to improve emission estimations and reduce energy consumption in the future [20].

## III. METHODOLOGY AND DATA

The method we used in this study is multilayer perceptron (MLP). It is a class of feedforward artificial neural networks (ANNs), which are best used for cases when input is high-dimensional and discrete, and the output is real-valued (human readability of the result is unimportant). MLP also reduces the negative impact of possibly noisy data [11]. The main reason we use MLP in this study is that our data might not be linearly separable.

Figure 1 represents the input eigenvector $X=\{x_1, x_2,\ldots, x_n\}$ and parameter vector $W = \{w_1, w_2,\ldots, w_n\}$. Compute it and get the cross product. The result is summed with the bias vector and input into the activation function Rectified Linear Unit (ReLU) to get the output. The above process is used to obtain neuron output results.

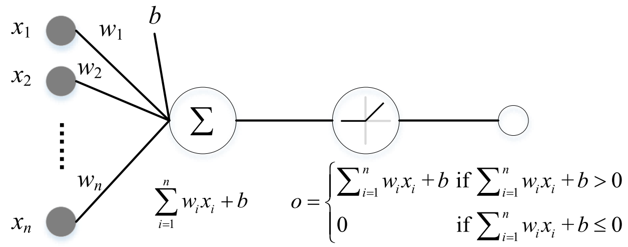

$x_1$
$x_2$
$x_n$
$w_1$
$w_2$
$w_n$
$b$
$\Sigma$
$\sum_{i=1}^{n} w_i x_i + b$
$o = \begin{cases} \sum_{i=1}^{n} w_i x_i + b & \text{if } \sum_{i=1}^{n} w_i x_i + b > 0 \\ 0 & \text{if } \sum_{i=1}^{n} w_i x_i + b \le 0 \end{cases}$

**FIGURE 1.** Artificial neural network perceptron architecture.

### A. THE ALGORITHM OF MULTI-LAYER PERCEPTION

Define the input as $\mathbf{X}_p = \{\mathbf{x}_{p1}, \mathbf{x}_{p2}, \ldots, \mathbf{x}_{pn}\}$, where $\mathbf{X}_p$ is the set of eigenvectors, $\mathbf{x}_{pn}$ is 417-dimensional eigenvector of the nth sample, and each element is a floating-point numerical feature. The corresponding tag is Y=$\{y_1, y_2, \ldots, y_n\}$, where Y is the set of tags, $y_n$ is the tag value of the nth sample and $y$ is a floating point value.

Generally, the number of layers of the network and the number of the neurons in each layer are empirical values. We compare the testing results with the Loss in the code, and finally decide to use four hidden layers in our multi-layer perceptron model, and the number of neurons in each layer is distributed as 20, 20, 50 and 50.

And the number of neurons in the hidden layer $l$ is $U_l$ and the output value is $\mathbf{O}_l$, so:

$$\boldsymbol{O}_l = f\left(\sum_{u=1}^{U_{l-1}} \boldsymbol{w}_l \boldsymbol{O}_{l-1} + \boldsymbol{b}_l\right) \tag{1}$$

where $\mathbf{w}_l$ is the weight matrix with dimension $U_l$ by $U_{l-1}$, $\mathbf{b}_l$ is the bias vector with dimension $U_l$. $f$ is a nonlinear activation function with ReLU used as the activation function, that is $f(\mathbf{O}_{l-1})$ =ReLU($\mathbf{O}_{l-1}$)

When $l$ =1, that is the input of the first hidden layer is the eigenvector of the data, that is

$$\boldsymbol{O}_{l=1} = f\left(\sum_{u=1}^{U_{l=0}} \boldsymbol{w}_{l=1} \boldsymbol{X}_p + \boldsymbol{b}_{l=1}\right) \tag{2}$$

Softmax is used as the output layer classifier, and the final predicted value is $\bar{y}$:

$$\bar{y} = S\left(O_{lout}\right) \tag{3}$$

where S is the Softmax function.

Divide the final fuel consumption into $k$ grades (a definite value should be found here), then the probability of obtaining grade $k$ is $\prod_k \bar{y}^{zk}$. In the training process, the z-label vector is one-hot real-valued, that is, only the kth dimension is 1, and the rest is 0 (it can be understood as 1 when classification is correct, otherwise 0). Finally, the loss function can be defined as $Loss = -\prod_{(x_p,z)} \prod_k \bar{y}^{zk}$:

$$Loss = -\prod\nolimits_{(x_p,z_n)} \prod\nolimits_k \bar{y}_n^{z_n k} = -\sum_n z_n \log \bar{y}_n \tag{4}$$

Equation (4) is the mean square error loss.

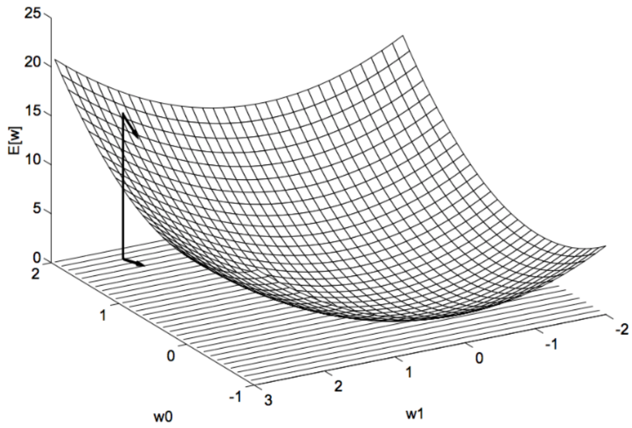


**FIGURE 2.** Relation between linear unit and squared error.

To better understand multiple classification, we first calculate binary classification, the loss function degenerates to

$$Loss'_{binary} = \bar{y}^z (1 - \bar{y})^{1-z} \tag{5}$$

The log likelihood function can be listed for all samples, and the loss function can be deduced as:

$$Loss_{binary} = \sum_{(x_p,z)} z \log \bar{y} + (1 - z) \log (1 - \bar{y}) \tag{6}$$

Equation (6) is the cross entropy loss.

This training rule guarantees that the output characteristics of MLP are linearly separable. The essential idea of MLP to solve the linear indivisibility is to map the original sample into the high-dimensional space through the kernel function, so that the sample is linearly separable in the high-dimensional feature space.

The linear unit training rule uses gradient descent, and it is guaranteed to converge to hypothesis with minimum squared error even when training data contain noises.

To explain the process of random gradient descent (shown in Figure 2), we specify the parametric optimization process.

Backward pass (BP) algorithm is used to deduce the gradient calculation process in multi-layer perceptron learning. As described in the first part, we can get the calculation of layer l in Equation (1).

According to the definition of Softmax:

$$\bar{y}_l = \frac{e^{\boldsymbol{O}_l}}{\sum_j e^{\boldsymbol{O}_j}} \tag{7}$$

Analyze $\bar{y}_{l'}$, and find the partial derivative of $\mathbf{O}_l$:

$$\frac{\partial \bar{y}_{l'}}{\partial \boldsymbol{O}_l} = \begin{cases} \dfrac{\sum_{l' \neq l} e^{\boldsymbol{O}_j} \cdot e^{\boldsymbol{O}_l}}{\sum_j e^{\boldsymbol{O}_j} \cdot \sum_j e^{\boldsymbol{O}_j}} = \bar{y}_l \left(1 - \bar{y}_l\right), & l' = l \\ -\dfrac{e^{\boldsymbol{O}_{l'}} \cdot e^{\boldsymbol{O}_l}}{\sum_j e^{\boldsymbol{O}_j} \cdot \sum_j e^{\boldsymbol{O}_j}} = -\bar{y}_{l'} \bar{y}_l, & l' \neq l \end{cases} \tag{8}$$

In which $\bar{y}_l = f(\boldsymbol{O}_l)$, Take the derivative of $\mathbf{O}_l$ for Loss, we have

$$\begin{aligned}
\frac{\partial Loss}{\partial \boldsymbol{O}_l} &= \frac{\partial\left[-\sum\limits_i z_i \log \bar{y}_i\right]}{\partial \boldsymbol{O}_l} \\
&= -\sum_i z_i \cdot \frac{\partial \log \bar{y}_i}{\partial \boldsymbol{O}_l} \\
&= -\sum_i z_i \cdot \frac{1}{\bar{y}_i} \cdot \frac{\partial \bar{y}_i}{\partial \boldsymbol{O}_l} \\
&= -z_l\left(1-\bar{y}_l\right) - \sum_{i \neq l} z_i \cdot \frac{1}{\bar{y}_i}\left(-\bar{y}_i \bar{y}_l\right) \\
&= -z_l + z_l \bar{y}_l + \sum_{i \neq l} z_i \bar{y}_l \\
&= -z_l + \bar{y}_l\left(\sum_i z_i\right) \\
&= \bar{y}_l - z_l
\end{aligned} \tag{9}$$

In which, $\sum\limits_i z_i = 1$.

At this point, the residual of the loss function on the optimal gradient is obtained, and the residual is reversely transferred back to the first hidden layer from the output layer, so as to realize the optimization of parameters of each layer along the gradient to the optimal direction.

Let $\delta_l = \frac{\partial Loss}{\partial \boldsymbol{O}_l}$, then the residuals are propagated layer by layer:

$$\begin{aligned}
\delta_l &= \frac{\partial Loss}{\partial \bar{y}_l} \cdot \frac{\partial \bar{y}_l}{\partial \boldsymbol{O}_l} = \frac{\partial \bar{y}_l}{\partial \boldsymbol{O}_l} \cdot \sum_{l-1} \frac{\partial \bar{y}_l}{\partial \bar{y}_{l-1}} \cdot \frac{\partial \bar{y}_{l-1}}{\partial \boldsymbol{O}_{l-1}} \\
&= Loss' \cdot \sum_{l-1} \boldsymbol{w}_l \delta_{l-1}
\end{aligned} \tag{10}$$

### B. DATA PROCESSING AND DATA FEATURE USAGE

The real-world fuel consumption data of light duty vehicles used in this paper comes from China's second ranked vehicle and traffic tool (BearOil APP, www.xiaoxiongyouhao.com). By the end of 2018, the APP had been downloaded 6 million times and has a monthly active user rate of 800 thousand. The accumulated mileage for all active vehicle owners from 31 different autonomous regions in China exceeds 23 billion kilometers, and the number of recorded data for real-world fuel consumption is over 51 million. The factors considered in this paper include host city, vehicle brand, vehicle type, engine parameter, gearbox type and fuel consumption provided by the MIIT. Because the APP also recorded the time each fuel consumption happened, we could use these data to regulate the difference from climate change and accordingly control the variability of the same vehicle based on which city the vehicle owner was in.

We use 2,424,379 samples to optimize our models. Among them, 484,875 samples are used as training data, 484,875 samples are used as validation samples and the rest are considered testing data. We use 20% of the total data as validation to prevent overfitting. If our parameters are fit on the training dataset, then this would result in a biased score. By including another validation dataset, we are able to lock away the test dataset while still being able to measure performance on unseen data as a way of selecting a good hypothesis. In our case, we use the training dataset for learning, fitting the parameters of the classifier and the validating the dataset for considering the number of hidden units in a neural network.

**TABLE 1. Dimension of original 8 features.**

| Features | Dimensions | Features | Dimensions |
|---|---|---|---|
| City | 17 | Engine type | 6 |
| Brand name | 106 | Engine capacity | 33 |
| Transmission model | 19 | Peak of engine output power | 193 |
| Month the data was generated | 5 | Reference fuel consumption | 14 |

This paper aims at predicting automobile fuel consumption, data adopted includes city, brand, vehicle type, engine model, transmission model, transmission type, gearbox shift form, reference peak of engine output power, engine displacement, reference fuel consumption, actual fuel consumption, time information of data produced, etc., which each data sample includes features.

The processing of the data includes defining the time range, that is, the time range generated by the selected data; removing the bad sample; obtain the corresponding environmental characteristics for the city where the sample data is generated and the corresponding time.

The eight features including city, brand name, engine model, transmission model, engine displacement, reference fuel consumption, peak of engine output power and the data generated are retained, that is, the feature vector of the original data has 8 dimensions. Through the processing of climate elements, a 24-dimensional feature vector is formed for each city, which is appended to the original 7-dimensional features to form a 31-dimensional feature.

Convert original 8 features into one hot representation, each feature dimensional is as Table 1:

Take actual fuel consumption as label, which means $y$ value is 1-dimensional.

Combined with the 24-dimensional climate features, excluding the label, each sample is represented by a 417-dimensional vector. Which means the $n$ in $\mathbf{X}_p = \{\mathbf{x}_{p1}, \mathbf{x}_{p2}, \ldots, \mathbf{x}_{pn}\}$ equals 417.

Take X as input and use Python to train, validate and test the dataset and import the 'sklearn' library to split the dataset and train inputs using MLP regression (see Figure 3). We then note the score, average, minimum, maximum and variance of minimum squared errors and the number of hidden layers used.

Additionally, we want to find the correlation between a certain factor and the result. For a single layer neural network, finding the correlation is simpler. Since the final result is a manipulation of all inputs, we could change each input one at a time and observe how the result changes. For example, if we have input $x$ and input $y$, and the result is calculated based on $x^2 + y$, then by picking a large $y$ value for the first trial and multiplying by 2 for the second trial, the result would differ by two times as well. However, if we try the same two values on $x$, then we can observe that the effect on the result is quadratic; therefore, we know the input $x$ should have more impact on the result than $y$ does. This is similar with MLP, only that now we have more inputs, more layers and at each layer, we apply operations to different inputs, so it is harder to observe the relation between certain inputs and the results.

## IV. RESULTS

### A. DESCRIPTIVE STATISTICAL RESULTS OF REAL-WORLD FUEL CONSUMPTION

Based on the BearOil database, we can see the difference of fuel consumption between MIIT and real-world fuel consumption (See Table 2).

From the results, we find that differences exist between real-world fuel consumption and standard fuel consumption released by MIIT. The actual fuel consumption is always higher than the standard consumption. Furthermore, there are obvious regional differences in real-world fuel consumption. The histogram of real-world fuel consumption also shows that the fuel consumption of most vehicle owners was less than 10. Although the standards in China fully meet the current level of the NEDC (New European Driving Cycle) standards for automotive oil products and several indexes even exceed NEDC standards, the NEDC standards may not be treated as the best reference for Chinese energy management. The NEDC standards are always inconsistent with the driving conditions in China. Based on the evaluation of the results from the working condition tests, it has been proven that the actual performance of vehicles in China is quite different from that in European countries. The China Automotive Testing Cycle (CATC) program is being launched by MIIT and other related ministries of China to design the standard of fuel consumption in China.

**TABLE 2. Descriptive statistical results of fuel consumption.**

| Variable | Obs | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|---|
| MIIT fuel consumption | 1,057,046 | 7.256 | 1.249 | 4.200 | 16.600 |
| Real-world fuel consumption | 1,057,046 | 9.071 | 2.359 | 3.000 | 19.997 |
| Shanghai | 216,611 | 9.487 | 2.431 | 3.003 | 19.997 |
| Lanzhou | 10,896 | 8.839 | 2.521 | 3.029 | 19.924 |
| Beijing | 147,335 | 9.193 | 2.339 | 3.000 | 19.987 |
| Nanning | 27,316 | 8.845 | 2.310 | 3.002 | 19.916 |
| Guangzhou | 117,228 | 8.891 | 2.352 | 3.001 | 19.992 |
| Kunming | 29,740 | 8.065 | 2.064 | 3.011 | 19.869 |
| Hangzhou | 61,932 | 9.252 | 2.331 | 3.023 | 19.992 |
| Wuhan | 84,567 | 9.044 | 2.234 | 3.044 | 19.971 |
| Shenyang | 67,017 | 9.150 | 2.441 | 3.006 | 19.992 |
| Jinan | 35,785 | 8.808 | 2.427 | 3.003 | 19.909 |
| Shijiazhuang | 31,907 | 8.433 | 2.160 | 3.052 | 19.979 |
| Fuzhou | 41,931 | 9.001 | 2.334 | 3.000 | 19.945 |
| Xining | 4,952 | 8.443 | 2.433 | 3.002 | 19.797 |
| Xi'an | 53,769 | 8.804 | 2.284 | 3.030 | 19.994 |
| Guiyang | 22,017 | 9.022 | 2.248 | 3.005 | 19.994 |
| Zhengzhou | 44,486 | 8.856 | 2.269 | 3.015 | 19.966 |
| Changsha | 59,557 | 9.047 | 2.265 | 3.011 | 19.980 |

### B. RESULTS OF MLP MODEL

Although stated in the methodology section, to evaluate the results of our MLP model, we present the output of our algorithm (See Table 3).

**TABLE 3. Result taken from MLP model.**

| Hyperparameter | MLP | MLP |
|---|---|---|
| Learning Rate | adaptive | adaptive |
| Hidden Units | 20 | 50 |
| Hidden Layers | 20 | 50 |
| Score | -2.495 | -5.404 |
| Average | 3,319.215 | 6,320.949 |
| Maximum | 3,498.277 | 6,600.915 |
| Minimum | 3,208.469 | 6,136.214 |
| Variance | 4,460.372 | 9,972.335 |

Within the model, we perform 10,000 iterations for the MLP regressor and calculate the minimum squared error for each iteration, and we analyze the performance of our model using the mean accuracy on the given test data and labels. A higher score indicates a better classification of the model; therefore, we pick the model with a lower number of hidden units and hidden layers. We also use the cross-validation method so that we can prevent overfitting the model with our training samples. If our model perfectly fits the training sample but fails to predict the validation sample, then it indicates that our model has a problem of overfitting. With cross validation, we first train using the training sample, then we use the validation sample to check its accuracy in prediction; then, we switch the role of the training sample and validation sample and repeat the above operation. In the end, we average the results. This gives us insight into how our model will generalize to an independent dataset.

To determine which factor influences the final fuel consumption prediction, we can alter the number of factors and see either when the algorithm converges or which results in the greatest decline in classification accuracy. Thus, this

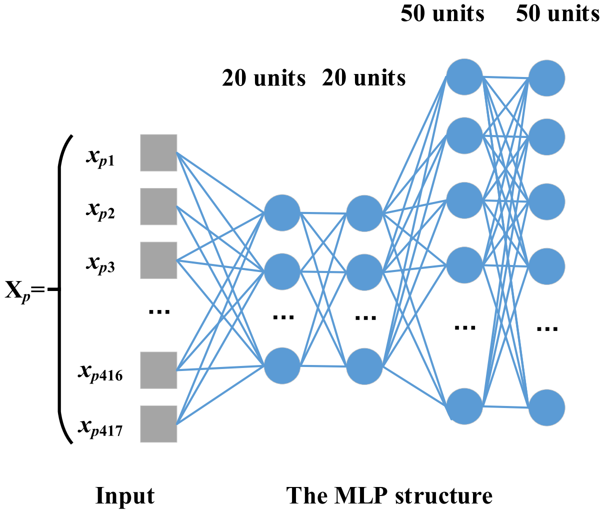


**FIGURE 3.** Procedure description of prediction.

process enables us to obtain a vague idea of which factor contributes most to the final result [12]. However, this approach is not precise because removing any of the inputs will result in a change in the NN architecture and its properties, and the complexity of the sigmoid functions should be considered.

There are other papers about fuel consumption prediction. One used the HDM-III model of the World Bank [7] and analyzed roughness levels of roads and engine characteristics to develop a unit model for fuel consumption. However, that study did not take into account as much data as we do and therefore did not represent as general a case as ours. Another paper used a logistic model to simulate the future trend of China's vehicle population [21]. However, although this prediction did tell us how much fuel consumption there might be in the future, it could not tell us how much consumption there is for each kind of vehicle with certain characteristics; it just gave us a national trend which is not accurate enough.

Since the MLP method provides accurate classification results despite the complicated property of different types of inputs, we use it for our fuel consumption prediction. Given the nature of MLP, it can be black box that gives little insight to its users of what is actually happening between the layers of hidden units, thus, utilizing sensitivity analysis could raise its transparency and the importance index returned can be used to help us identify which factor affects fuel consumption the most.

### C. COMPARISION OF PREDICTED RESULTS WITH ACTUAL VALUE

We divide the data set into training set and test set, in which the number of training set samples accounts for 70% and the number of test sets accounts for 30%, in which there is no intersection between training set and test set. On the basis of mini-batch and dropout theory, supervised training is conducted on the involved models through the training set, and the models are evaluated for the test set after convergence.

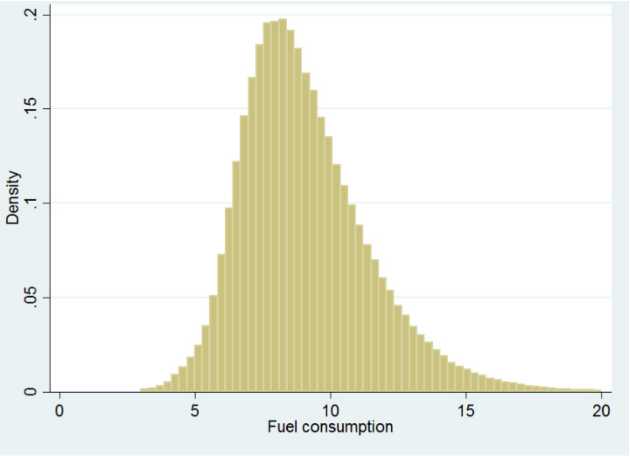


**FIGURE 4.** Histogram of real-world fuel consumption.

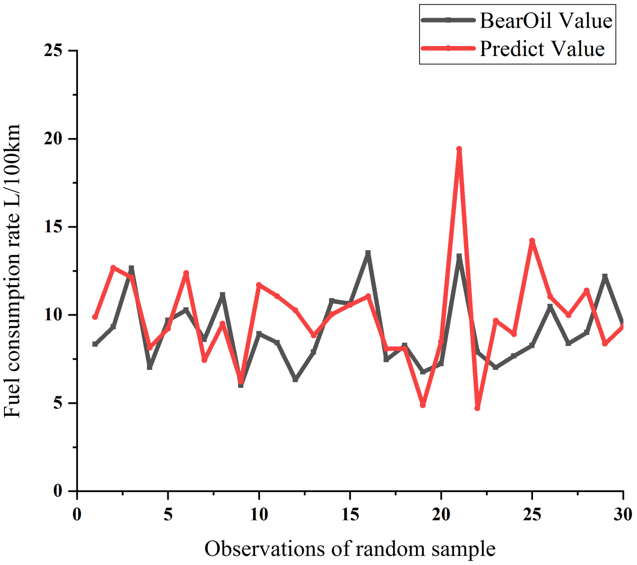


**FIGURE 5.** Distribution of predicted results.

We randomly select 30 test results from the test set samples for display, and the predicted results are shown in Figure 5.

As shown in the Figure 5, the predicted value and the real value of the randomly selected samples are generally consistent with the trend of sample change. Affected by real factors, the method to make predictions for certain samples still contains certain errors, as shown in figure, the predicted peak and valley space corresponding to the estimated values still contains some errors compared to the actual. The reason for this is that this method, compared with traditional regression prediction method, considers a lot more factors, and is hard to avoid error on weights of some secondary factors, but from the trend of the overall point of view, the effectiveness of the chosen model is proven.

In addition, on the basis of the same model parameters, we filter out subset of automatic transmission (AT) models and subset of manual transmission (MT) models based on transmission model in the data parameters, and then per-

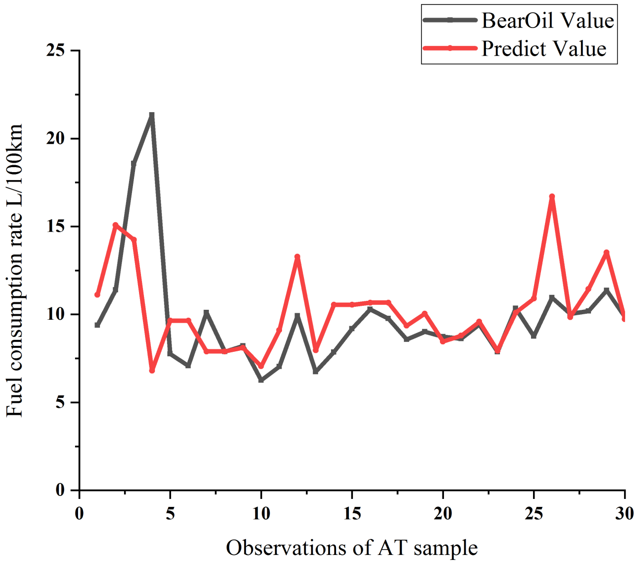


**FIGURE 6.** Distribution of automatic transmission vehicle.

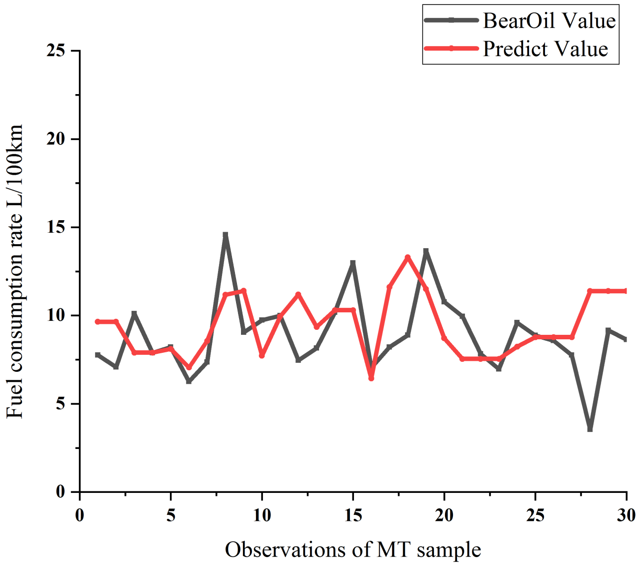


**FIGURE 7.** Distribution of manual transmission vehicle.

form fuel consumption prediction on these two subsets. Fuel consumption distribution of actual values and predicted values of the AT vehicle is shown in Figure 6, distribution of MT vehicle is shown in Figure 7.

As shown in Figure 6 and Figure 7, the model adopted in this paper is more consistent with the change trend of the real values in terms of the predicted value of fuel consumption of the automatic transmission model compared with the direct sense analysis.

## V. CONCLUSIONS

In this paper, based on a real-world database that is originally collated for the purpose of fuel consumption prediction, we demonstrate an approach for exploring the main influencing factor on fuel consumption prediction. The model in our study includes the parameters of external environmental factors, the manipulation of vehicle companies and the drivers' driving habits.

Combined with sensitivity analysis, we find that using MLP best classifies the given nonlinear dataset and that the architectures are capable of learning powerful features. To further improve our method, it is easier to split the dataset, other than the test dataset, into more sections to perform a more accurate cross validation. However, such an approach requires our dataset to be larger than it already is, and it requires time for data accumulation.

Additionally, a more precise way of analyzing the importance of each factor for an MLP model is required. As the sensitivity analysis currently only improves the transparency of an MLP model, it is still not able to give us a clear view of how any inputs actually affect the output. In the future, we hope to optimize our model and obtain a better understanding of the relationship between the input factor and output results.

## ACKNOWLEDGMENT

The authors thank the founders of the BearOil APP (www.xiaoxiongyouhao.com), Xiong Zhang and Hongbo Sun, for providing access to their company's unique fuel consumption rate of light duty vehicles data.

## REFERENCES

[1] M. André, R. Joumard, R. Vidon, P. Tassel, and P. Perret, "Real-world European driving cycles, for measuring pollutant emissions from high- and low-powered cars," *Atmos. Environ.*, vol. 40, no. 31, pp. 5944–5953, 2006.

[2] M. B. Arias and S. Bae, "Electric vehicle charging demand forecasting model based on big data technologies," *Appl. Energy*, vol. 183, pp. 327–339, Dec. 2016.

[3] H. Cai, X. Jia, A. S. F. Chiu, X. Hu, and M. Xu, "Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet," *Transp. Res. D, Transp. Environ.*, vol. 33, pp. 39–46, Dec. 2014.

[4] N. Davis, J. Lents, M. Osses, N. Nikkila, and M. Barth, "Development and application of an international vehicle emissions model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1939, no. 1, pp. 156–165, 2005.

[5] M. De Gennaro, E. Paffumi, and G. Martini, "Big data for supporting low-carbon road transport policies in Europe: Applications, challenges and opportunities," *Big Data Res.*, vol. 6, pp. 11–25, Dec. 2016.

[6] H. Guo, Q.-Y. Zhang, Y. Shi, and D.-H. Wang, "Evaluation of the international vehicle emission (IVE) model with on-road remote sensing measurements," *J. Environ. Sci.*, vol. 19, no. 7, pp. 818–826, 2007.

[7] X. Li and Z. Yao, "Vehicle fuel consumption prediction model," *J. Tongji Univ.*, (in Chinese), vol. 20, no. 4, pp. 403–410, 1992.

[8] Y. Li, W. Jiang, L. Yang, and T. Wu, "On neural networks and learning systems for business computing," *Neurocomputing*, vol. 275, pp. 1150–1159, Jan. 2018.

[9] Y. Li, L. Yang, B. Yang, N. Wang, and T. Wu, "Application of interpretable machine learning models for the intelligent decision," *Neurocomputing*, vol. 333, pp. 273–283, Mar. 2019.

[10] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[11] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

[12] Z. Ma, Y. Lai, W. B. Kleijn, Y. Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.

[13] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.

[14] T. M. I. Mahlia, S. Tohno, and T. Tezuka, "A review on fuel economy test procedure for automobiles: Implementation possibilities in Malaysia and lessons for other countries," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 4029–4046, 2012.

[15] H. Rahimi-Eichi and M.-Y. Chow, "Big-data framework for electric vehicle range estimation," in *Proc. 40th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct./Nov. 2014, pp. 5628–5634.

[16] C. Silva, M. Ross, and T. Farias, "Evaluation of energy consumption, emissions and cost of plug-in hybrid vehicles," *Energy Convers. Manage.*, vol. 50, no. 7, pp. 1635–1643, 2009.

[17] Q. Shi and M. Abdel-Aty, "Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 380–394, Sep. 2015.

[18] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 162–177, Sep. 2015.

[19] D. V. Wagner, F. An, and C. Wang, "Structure and impacts of fuel economy standards for passenger cars in China," *Energy Policy*, vol. 37, no. 10, pp. 3803–3811, 2009.

[20] Q. Wang, H. Huo, K. He, Z. Yao, and Q. Zhang, "Characterization of vehicle driving patterns and development of driving cycles in Chinese cities," *Transp. Res. D, Transp. Environ.*, vol. 13, no. 5, pp. 289–297, 2008.

[21] Y. Wu, P. Zhao, H. Zhang, Y. Wang, and G. Mao, "Assessment for fuel consumption and exhaust emissions of China's vehicles: Future trends and policy implications," *Sci. World J.*, vol. 2012, pp. 1–8, Nov. 2012.

[22] T. Zachariadis, "On the baseline evolution of automobile fuel economy in Europe," *Energy Policy*, vol. 34, no. 14, pp. 1773–1785, 2006.

[23] X. Zheng *et al.*, "Big data for social transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2015.

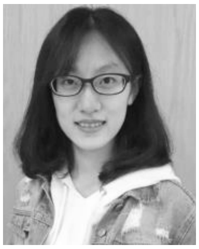
[24] K. Zhou, S. Yang, C. Shen, S. Ding, and C. Sun, "Energy conservation and emission reduction of China's electric power industry," *Renew. Sustain. Energy Rev.*, vol. 45, pp. 10–19, May 2015.

**YAWEN LI** received the Ph.D. degree in innovation, entrepreneurship, and strategy from Tsinghua University, in 2018. She is currently an Assistant Professor with the School of Economics and Management, Beijing University of Posts and Telecommunications. Her research papers have been published in or accepted by journals, including the *Journal of Cleaner Production*, *Neurocomputing*, the *Asian Journal of Technology Innovation*, the *Journal of Leadership and Organizational Studies*, and *Chinese Medical Journal*. Her research interests include the collaborative innovation, the development of science parks, and scientific productivity of firms.

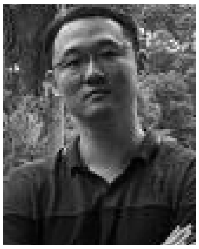
**GUANGCAN TANG** is currently a Research Assistant with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His research interests include financial technology, machine learning, and computer vision.

**JIAMENG DU** received the degree in electrical and computer engineering with Carnegie Mellon University. Her research interests include artificial intelligence, machine learning, and data analysis.

**NAN ZHOU** is currently pursuing the Ph.D. degree in computer science with the Beijing University of Posts and Telecommunications. His research interests include deep learning, information retrieval, and machine learning.



**YUE ZHAO** is currently pursuing the degree with the International School, Beijing University of Posts and Telecommunications. Her research interests include business analytics and data analysis.



**TIAN WU** received the Ph.D. degree in finance from Tsinghua University, in 2016. He is currently an Assistant Professor with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research papers have been published in journals, including *Energy Policy*, *Energy*, the *Journal of Cleaner Production*, *Transportation Research Part A: Policy and Practice*, *Neurocomputing*, *Resources, Conservation & Recycling*, the *International Journal of Environmental Research and Public Health, Sustainability, Energies*. His current research interests include operations management, industrial organization, sharing economy, innovation, entrepreneurship, and strategy.