Type of the Paper (Article)

Estimation of Real-world Fuel Consumption Rate of Light-duty Vehicles Based on Big Data

Isabella Yunfei Zeng [1], Shiqi Tan [2], Jianliang Xiong [3,*], Xuesong Ding [4], Yawen Li [5], and Tian Wu [6]

[1] UK-China (Guangdong) CCUS Centre; isabellazeng04@gmail.com
[2] Department of Automation, Tsinghua University; tsq19@mails.tsinghua.edu.cn
[3] School of Economics and Management, Tsinghua University; xjl19@mails.tsinghua.edu.cn
[4] School of Economics and Management, Beijing University of Posts and Telecommunications; dingxuesong2000@bupt.edu.cn
[5] School of Economics and Management, Beijing University of Posts and Telecommunications; warmly0716@bupt.edu.cn
[6] Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences; wutian@amss.ac.cn

* Correspondence: xjl19@mails.tsinghua.edu.cn

Abstract: Private vehicle travel is the most basic mode of transportation, and the effective control of the real-world fuel consumption rate of light-duty vehicles plays a vital role in promoting sustainable economic development as well as achieving a green low-carbon society. Therefore, the impact factors of individual carbon emission must be elucidated. This study builds five different models to estimate real-world fuel consumption rate of light-duty vehicles in China. The results reveal that the Light Gradient Boosting Machine (LightGBM) model performs better than the linear regression, Naïve Bayes regression, Neural Network regression, and Decision Tree regression models, with mean absolute error of 0.911 L/100 km, mean absolute percentage error of 10.4%, mean square error of 1.536, and R squared (R2) of 0.642. This study also assesses a large number of factors, from which three most important factors are extracted, namely, reference fuel consumption rate value, engine power and light-duty vehicle brand. Furthermore, a comparative analysis reveals that the vehicle factors with greater impact on real-world fuel consumption rate are vehicle brand, engine power, and engine displacement. Average air pressure, average temperature, and sunshine time are the three most important climate factors.

Keywords: Real-world fuel consumption rate; machine learning; big data; light-duty vehicle; China

1. Introduction

Tightening the oil consumption has always been one of the focuses of building a greener city, including the limitation of gasoline. Recently, a new round of investigation of fine particle sources in Beijing was officially released. The results reveal that coal combustion is no longer the main source of PM2.5 in Beijing, and mobile sources such as vehicles have become the primary source of pollutants. To date, China has implemented a series of measures to control the fuel consumption rate of vehicles. In September 2019, Ministry of Industry and Information Technology (MIIT) of the People's Republic of China and other relevant ministries issued the "Decision on Amending the Measures for the Parallel Management of Average Fuel Consumption of Automobile Enterprises and New Energy Vehicle Score". The objective of the automobile enterprise fuel consumption score is to promote the sustainable development of China's new energy vehicle industry, accelerate the transformation of energy structure, upgrade the traditional gasoline vehicle industry, and achieve a set of other goals in accordance with carbon neutrality. To improve the performance and accuracy of fuel consumption score, in which aims at reducing fuel consumption, the most effective method is to increase the production of pure electric and plug-in hybrid electric vehicles.

 doi:10.20944/preprints202111.0029.v1

The current fuel consumption score is calculated from the fuel consumption by MIIT, which can be roughly divided into the following steps. The first step is to calculate the average fuel consumption of each automobile enterprise according to the national standard (GB27999-2014). This is calculated on the basis of the weighted average of the output of each vehicle and the fuel consumption value specified in the standard. The fuel consumption of each vehicle is closely related to the vehicle's curb weight, but the curb weight varies significantly among different vehicles. Therefore, the required fuel consumption standard are also different. The second step is to calculate the fuel consumption reported by each automobile enterprise for the corresponding vehicle types according to MIIT. The third step is to calculate the difference between the 2018 standard and the 2018 actual fuel consumption (the fuel consumption reported by MIIT) multiplied by the output, which is exactly the fuel consumption score for the specific vehicle enterprise.

The "Limits and measurement methods for emissions from light-duty vehicles (CHINA 6)" guidelines, which are issued jointly by Ministry of Ecology and Environment and the General Administration of Quality Supervision, Inspection and Quarantine, require that all sold and registered light vehicles satisfy the standards, starting from July 1, 2020. According to "Energy Conservation and New Energy Automobile Industry Development Plan (2012–2020)", the average fuel consumption rate of passenger vehicles in China should be reduced to 5.0L/100 km by 2020. MIIT has promulgated the "Measures for Parallel Management of Average Fuel Consumption and New Energy Vehicle Integral in Passenger Vehicle Enterprises", which was implemented on April 1, 2018. The promulgation and implementation of these policies impose higher requirements for energy saving and emission reduction technology in the automobile industry. To solve the current energy and environmental problems and achieve carbon neutrality in the near future, it is very important to estimate the real-world fuel consumption of light-duty vehicles and determine its impact factors.

To date, the most direct approach to determining the fuel consumption rate of a vehicle is to check the reference fuel consumption information provided by MIIT, which may be far different from the actual fuel consumption. Since the implementation of vehicle emission test standard, China has been adopting the New European Driving Cycle (NEDC) working conditions to test fuel consumption and emissions. However, some problems have arisen after years of practice. The NEDC working condition test results are quite different to the real-world driving situation in China, which not only interferes with the judgement in terms of understanding the actual driving state, but also does harm to government credit from the perspective of vehicle drivers.

The problem of the NEDC working condition is mainly manifested in three aspects. First, the NEDC working condition is very different to the driving characteristics of automobile vehicles in China. This difference is particularly evident in the emission performance, fuel consumption, and optimized calibration value based on the NEDC. Second, this divergence directly affects the implementation of China's energy conservation and emission reduction policies, which has a negative impact on the government reputation to some extent. Third, the existing NEDC working condition method underestimates the energy saving effect of new energy vehicles.

In fact, the NEDC condition is too ideal in three main aspects. First, there is a large difference between the laboratory simulation conditions and the actual road conditions. Specifically, China has a vast territory and the road conditions in different regions greatly vary, which is neglected by NEDC condition. Second, the NEDC working condition test ignores the influence of external factors, such as air pressure and temperature, which may influence the fuel consumption to a certain degree. Third, the NEDC working condition test does not consider the actual behavior of the driver, such as their driving habits and use of care air conditioner.

Hence, the China Automotive Test Cycle (CATC) was launched in 2015. Compared with the NEDC (European Fuel Consumption and Emissions Assessment Standard) adopted by the National Five Emission Standards, and the World Light Vehicle Testing

 doi:10.20944/preprints202111.0029.v1

Regulations (WLTC) working conditions adopted by the National Sixth Emission Standard, CATC's working conditions more realistically reflect the actual condition of China's roads. The successful introduction of this project enable an independent basic standard system for the Chinese auto industry.

Three different types of data are collected with regard to China's working conditions. First, the collection of real-time and synchronized large-scale driving data for different vehicles in different regions is realized using CAN+GPRS technology. Second, geographic information system all-road low-frequency dynamic big data are used to calculate the actual turnover of the vehicle at different speed intervals and its coefficient, which reflect the macroscopic distribution of the vehicle in different speed-ranges in a more objective and accurate manner. Third, the driving behavior characteristics, air-conditioner usage characteristics, and other characteristics of the vehicle are investigated.

The data used in this study were obtained from the BearOil app (www.xiaoxiongyouhao.com), which has already been downloaded 6 million times with more than 800 thousand active monthly users. The accumulated mileage of active vehicles in 31 different provincial regions of China has exceeded 23 billion kilometers, and the real-world fuel consumption rate records have exceeded 51 million. Moreover, this study also takes vehicle factors such as vehicle brand, engine power, engine displacement, as well as climate and environmental factors such as average air pressure, average temperature, and sunlight hours into consideration. Therefore, this study aims at discovering the most important factors with impact on the real-world fuel consumption rate of vehicles.

The rest of the paper is organized as follows. First, the related literature is reviewed in Section 2. The data source, extracted real-world fuel consumption rate and climate factors are discussed in Section 3. Section 4 describes the experiments, including model selection and model training. Besides, Section 4 also reports the results, including the comparison of different models and assessment of the most important features. Section 5 discusses the feature importance assessment. Section 6 presents the conclusions and the implications with regard to policy.

2. Literature review

Considering the large proportion of environmental pollution that could be accounted for automobile source [1, 2], it is important to obtain relatively accurate fuel consumption information. Furthermore, the application of artificial intelligence in business intelligence rises gradually [3]. Therefore, models for estimating the real-world fuel consumption rate and assessment of impact factors are being proposed at an increasing pace.

Li et al. [4] used a multilayer perceptron (MLP) method to estimate the fuel consumption rate of light-duty vehicles. Their model considered parameters including external environmental factors, the manipulation of vehicle companies, and the driving habits of drivers. It was found that multilayer perceptron method could classify their nonlinear dataset in the most reasonable way under sensitivity analysis. However, the sensitivity analysis only increased the MLP model's transparency, but did not analyze the importance of each factor in the MLP model or elucidate the way in which the outputs change with different inputs. Some studies have used a two-level clustering model to determine the driving patterns of electric vehicles. These studies extracted the driving pattern characteristics, namely, the mileage range and parking range from an electric vehicle dataset. Then, the driving patterns, which were daily driving patterns and multifaceted driving patterns, were estimated using a two-level clustering model. Yet this model only focused on simple vehicle static parking patterns and did not consider other traffic information and the weather conditions [5]. Wu et al. [6] predicted the fuel consumption rate by learning from real-world data of vehicle owners.

Although many models have been proposed to estimate fuel consumption, the HDM-4 fuel consumption model has been widely used in most cases. Many studies have used the HDM-4 fuel consumption model and then carried out calibration, which is a necessary step in this methodology [7, 8]. The accuracy of the HDM-4 fuel consumption model and the need for further calibration were discussed in [9]. This study was based on a limited

set of tests, wherein a small number of vehicles were tested at constant speed on selected sections under limited weather conditions. Therefore, it is not clear whether these estimates reflect the actual fuel consumption under realistic driving conditions. Additionally, the authors proposed that the vehicle weight and frontal area should be given more consideration. Yamashita et al. [10] developed a forecasting model based on the driving behavior to estimate fuel consumption. They used Pearson coefficient correlation analysis based on data mining to filter the driving behavior indicators, which were highly correlated with fuel consumption. The highly correlated driving behavior indicators used in the model can be classified into four categories: speed, acceleration, Left/Right/U-turn, and other indicators. Through neural network modeling and regression analysis, these highly correlated driving behavior indicators generated more than 12 aggregation models. Moreover, the best mean absolute percentage error value among them was below 5%. These categories and mean absolute percentage error provide a certain reference for the assessment of driving behavior. Ahn et al. [11] used the microscopic fuel consumption and emission model to predict fuel consumption of normal light-duty vehicles based on the instantaneous vehicle speed and acceleration levels. It was found that the vehicle emissions generated by these models are consistent (in excess of 90%) with the measured coefficients in the Oak Ridge National Laboratory data. The authors attempted to develop these models to bridge the gap between existing traffic simulation models, traditional transportation planning models, and environmental impact models. But their models could be improved to expand applicability, and environmental factors, impact of heavy-duty vehicles on the environment, and high-emitting vehicles should be further considered in such models.

Besides, models can also be used to estimate the vehicle fuel consumption and emissions directly through instantaneous Global Positioning System (GPS) speed measurements [12], and succeeded in extensively assessing the efficiency, energy, environmental, and safety benefits. Specifically, the assessment included the counts of the evaluation of the midblock tube, number of the intersection turning, speed measurement per second from GPS-equipped floating cars, and evaluation of the traffic signal coordination network. However, this study only reported field evaluation results. Lei et al. [13] proposed the Microscopic Emission and Fuel consumption model for two categories of light-duty vehicles, which are widely used and have been shown to be effective. Compound acceleration variables were introduced into the Microscopic Emission and Fuel consumption model to capture the effects of the interaction between the historical acceleration and current speed on emissions and fuel consumption. After calibration, the instantaneous verification results reveal that the Microscopic Emission and Fuel consumption model performed better when the mean absolute percentage error was lower than that of the other two models. Additionally, the overall verification results reveal that the Microscopic Emission and Fuel consumption model produced reasonable estimates compared with the actual measurements.

Another study used a vehicle-specific fuel consumption model based on a PEMS application to estimate fuel consumption under different driving patterns. The vehicle fuel consumption per unit time exhibited strong positive correlation with the cruise speed. The fuel consumption rate appeared to be optimal in the speed range of 50–70 km/h. When the vehicle accelerated, the fuel consumption rate significantly increased, but only slightly changed when the vehicle decelerated. In each of these speed categories, linear functions and exponential functions were derived for the fuel consumption rates and vehicle specific power bins, respectively. The travel fuel consumption and fuel consumption rate generated by the vehicle specific power-based model were accurate to approximately +15% and +20% [14].

Existing studies have shown that real-world fuel consumption rate is influenced by the objective characteristics of the road, such as the road surface [15, 16], road width [17, 18], traffic congestion and speed limits [19, 20], energy management strategy [21-23], and fuel tank status monitoring technology [24]. Ejsmont [16] handled the above-mentioned

 doi:10.20944/preprints202111.0029.v1

factors by investigating the relationship between the surface texture and the rolling resistance of light and heavy vehicle types. He used the mean profile depth as a parameter to proxy the road surface, which is correlated with the rolling resistance of different vehicle types. Additionally, the performance results reveal that, although correlation exists, it cannot be explained in absolute terms because the regression between the mean profile depth and the rolling resistance is not linear. Kono [17] considered many factors, including traffic information, geographic information, vehicle parameters, and driver behavior, to analyze and predict fuel consumption. The author proposed a fuel consumption prediction model for ecological route search and compared its results to that obtained by the traditional time priority route search method and a driving experiment. The author concluded that it is important to propose an indicator of fuel reduction effectiveness for future emission reduction technologies, including ecological route search. Brundell-Freij [19] reported that speed and other factors, such as the acceleration and type of gears, influence fuel consumption. His study aimed at better understanding the variables affecting the driving patterns by determining the impact of street characteristics and driving habitat. He analyzed the relationship between the driving patterns and the potential variables, and found that the most obvious factor involved the travelling speed limit, whose impact was more relevant at ninety kilometers per hour or lower. Finally, the results reveal that the influence of the street and traffic environment on the driving behavior is dependent on driver variables and vehicle performance.

Additionally, real-world fuel consumption rate is affected by climate. For example, winter has been related to a decrease of 20% in fuel efficiency [21]. Other studies have established the relationship between temperature and driving environment [22, 23]. Zahabi [21] investigated fuel efficiency, and then compared vehicle performance to that of a standard gasoline vehicle in a cold Canadian urban environment. He considered many different factors including the driving conditions, temperature, and speed. In his results, low temperature below 0 °C in winter, was identified as a factor exerting detrimental influence on fuel consumption. Specifically, it was found that fuel efficiency decreased by 20% in winter compared with that in summer. In the present study, the climate environment is also an important factor and the temperature factor is discussed in detail. Weilenmann, Favez, and Alvarez [22] proposed that cold starting, which refers to the internal temperature of vehicles, can reduce the emission of modern gasoline and diesel passenger cars. Alvarez and Weilenmann [23] proposed that low ambient temperatures affect hybrid electric vehicles in terms of fuel consumption, and investigated these characteristics in five in-use hybrid electric vehicle models.

Subjective characteristics, such as driving velocity [19, 24] and driving acceleration [25], also affect real-world fuel consumption rate and are used to describe the temporal characteristics of driving patterns. Generally, existing studies have mainly focused on the actual road conditions, environmental factors, and driving behavior, but did not rank importance priority to these factors. Xu, Chen, and Li [24] reported that speed has a remarkable effect on fuel consumption, particularly when the vehicles travel on urban roads where there are many traffic signals. Hence, to reduce fuel consumption, the authors proposed a double-layer speed optimization method with real-time computation, and obtained the optimal real-time, which demonstrates the potential of the double-layer speed optimization method in improving fuel consumption and reducing travel time. Wang [25] analyzed the driving characteristics and established driving cycles. By comparing Chinese cities with European and American cities, this study concluded that the average speed, average acceleration, and percentage of acceleration time are different in these different regions.

Our research proposes five models, namely, the linear regression, Naïve Bayes regression, Neural Network regression, Decision Tree regression, and LightGBM models, to estimate real-world fuel consumption rate of light-duty vehicles in China. The results obtained by these five models are compared to determine the optimal one. Additionally, this study assesses 17 different factors and ranks the importance priority of each factor.

3. Materials and Methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

3.1. Data

The data used in this study were obtained from two sources: the real-world fuel consumption rate records reported by vehicle owners in the BearOil app, and the monthly dataset of the surface climate in some regions of China.

3.1.1. Fuel consumption rate information

In this study, about 2 million records of real-world fuel consumption rates reported by vehicle owners in 17 provincial capitals of China in the period of 2013–2017 were extracted from the BearOil app. Examples of the real-world fuel consumption rate data are shown in Table 1. To protect user privacy, the user number (User_ID) only shows the last eight digits of the true value.

Table 1. Raw data example of real-world fuel consumption rate information from BearOil APP.

Feature Name	Instance 0	Instance 1	Instance 2	Instance 3	…
User_ID	65961294	17424034	28206249	78105203	…
City	Hangzhou	Shanghai	Wuhan	Guangzhou	…
Date	2017/06/28	2013/06/26	2013/07/24	2019/06/26	…
Brand Name	BMW	ROEWE	SKODA	TOYOTA	…
Series Name	BMW X1	ROEWE 350	FABIA	LEVIN	…
Version Year	2016	2011	2011	2016	…
Engine	1.5L/136ps/L3	1.5L/109ps/L4	1.4L/86ps/L4	1.8L/99ps/L4	…
Gearbox	AMT-6	MT-5	MT-5	E-CVT	…
Refconsumption (L/100km)	6.1	7.8	6.5	4.2	…
Consumption (L/100km)	11.8	9.6	7.0	4.6	…

The User_ID in the sample is the unique ID of a BearOil app user. Therefore, the same User_ID corresponds to several samples and was used to record the time-varying relationship between the user's real-world fuel consumption rate, including the reporting time and the city in which the user lives, and the fuel consumption rate measured by the user.

The relevant information of the vehicle is given in the sample, including the vehicle brand , series and versions. Because the exact version of different vehicles brands is quite different, therefore only the version year of the each example is shown here. Additionally, the sample features include information of the vehicle engine and transmission. The engine parameters include the displacement (unit: L), power (unit: ps), and cylinder number. The transmission parameters indicate the type of transmission, including manual transmission (MT), automatic transmission (AT), automated manual transmission (AMT), continuously variable transmission (CVT), direct shift gearbox (DSG), and so on.

Additionally, our dataset also includes a reference value for fuel consumption rate of the corresponding vehicle, which is provided by MIIT of China. The fuel consumption rate measurement method adopted by MIIT refers to the second stage of NEDC. However, there exists various problems, such as incompatibility with the current vehicle power and the overall quality, and a great gap between the actual driving conditions. Besides, owing to the impact of different climate conditions, driving behaviors, and other factors, the reference fuel consumption rate often poorly proxies real-world fuel consumption rate.

Moreover, because some information is often omitted by APP users in the process of data uploading, there are many missing values in the original dataset. The corresponding processing methods are introduced in the data preprocessing section of this paper.

3.1.2. Climate information

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

The climate information data were extracted from the Monthly Report of Surface Meteorological Observation provided by meteorological departments of the provincial regions in China. In this study, the climate data from 2013 to 2017 were used, which is consistent with the spatial range of fuel consumption rate data.

Each climate data contains the station number of the climate observation area, and annual and monthly statistical information. The relevant climate characteristics, specific meanings, and units of measurement are listed in Table 2.

Table 2. Factors of meteorological data and unit of measurement.

Feature number	Feature Name	Unit
V10004	Average pressure	0.1hPa
V10201	Extreme maximum pressure	0.1hPa
V10202	Extreme minimum pressure	0.1hPa
V13004	Mean vapor pressure	0.1hPa
V12001	Average temperature	0.1℃
V12011	Extreme maximum temperature	0.1℃
V12012	Extreme minimum temperature	0.1℃
V12211	Mean maximum temperature	0.1℃
V12212	Mean minimum temperature	0.1℃
V12201	Average temperature anomaly	0.1℃
V13003	Mean relative humidity	1%
V13007	Minimum relative humidity	1%
V11002	Average wind speed	0.1m/s
V11042	Maximum wind speed	0.1m/s
V11041	Extreme maximum wind speed	0.1m/s
V11212	Maximum wind direction	azimuth
V11043	Extreme maximum wind direction	azimuth
V13011	Average precipitation	0.1mm
V13052	Maximum daily precipitation	0.1mm
V13212	Precipitation anomaly percentage	1%
V13012	Daily precipitation >=0.1mm days	1day
V14033	Sunshine percentage	1%
V14032	Sunshine time	0.1h

As can be seen, the climate information includes the temperature, barometric pressure, precipitation, sunlight, and other information. The climate information of different regions during the sample period also exhibits great variation, which has a non negligible impact on real-world fuel consumption rate of automobiles.

Because the climate of a certain region exhibits regularity within a certain month, this study treated the average climate condition in different cities and different months as the climate factors. As for wind direction, the north wind is defined as 1, and this number increases by 1 every 22.5 degrees clockwise. Additionally, if the wind speed is less than or equal to 0.2 m/s, the wind is considered to be calm, which corresponds to number 17. Therefore, there are totally 17 wind direction categories successively numbered from 1 to 17.

3.2. Factor Extraction

3.2.1. Factor extraction of fuel consumption rate information

The fuel consumption rate information obtained from the BearOil app mainly includes three factors: the vehicle factors, reference fuel consumption rate, and real-world fuel consumption rate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

First, the objective of this study should be clarified. For a certain user who drives the same car, there is a certain fluctuation in the fuel consumption value reported each time, which is attributed to differences in the driving behavior and driving environment at different times. This study aimed to predict the average real-world fuel consumption rate of specific vehicle types under specific climate conditions. Therefore, the real-world fuel consumption rate reported by a specific APP user in different cities and months was averaged and treated as the prediction target.

There are significant differences among fuel consumption rates for different vehicle brands, engine parameters, and transmission parameters. Therefore, this study selected the above factors as the model input. The displacement and power characteristics of the engine parameters are continuous variables, while the other characteristics are discrete variables. Because the number of vehicle series belonging to different brands are too large in our data set, there will be too many dimensions if we employ one-hot encoding. Since a certain correlation exists between the proposed parameters and exact vehicle series, the vehicle series are not used as input.

Moreover, although many studies have reported that the reference value by MIIT and the actual fuel consumption rate are quite different, these reported official data can still act as a reasonable range of the vehicles' actual fuel consumption rate and can also be used as a reference for eliminating abnormal fuel consumption values uploaded by the APP users. This study, therefore, includes the reference consumption by MIIT which is a continuous variable, as an input feature.

3.2.2. Factor extraction of climate information

The available climate factors are listed in Table 3. We merge the fuel consumption information with the corresponding climate information in different cities and dates, which are combined to be input variables in our models. Additionally, to prevent multicollinearity originated from strong correlation between the climate characteristics, it was necessary to test the correlation coefficient between these input variables. The climate variable pairs with correlation coefficients above 0.8 are listed in the Table 3.

Table 3. Correlation table of climate factors.

Feature A	Feature B	Pearson correlation
V12012	V12212	0.99296
V12012	V12211	0.96847
V12012	V12001	0.98527
V12012	V13007	0.70025
V12012	V13004	0.95931
V12012	V12011	0.93381
V14032	V14033	0.90305
V12212	V12211	0.975585
V12212	V12001	0.99426
V12212	V13004	0.95482
V12212	V12011	0.95523
V13212	V13012	0.99583
V10202	V10004	0.99927
V10202	V10201	0.99769
V10004	V10201	0.99944
V12211	V12001	0.99237
V12211	V13004	0.93057
V12211	V12011	0.97633
V11042	V11041	0.83803
V12001	V13004	0.94535
V12001	V12011	0.97182
V13007	V13003	0.86545

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

V13004	V12011	0.88903

As can be seen, there is strong correlation between many climate-related variables, which require us to select a proper set of corresponding characteristics. For each variable pair with strong correlation, only one characteristic is selected, and all the selected input characteristics of climate factors are listed in Table 4.

Table 4. Selected climate factors.

Feature number	Factor Name	Unit
V10004	Average pressure	0.1hPa
V12001	Average temperature	0.1℃
V12201	Average temperature anomaly	0.1℃
V13003	Mean relative humidity	1%
V11002	Average wind speed	0.1m/s
V11212	Maximum wind direction	azimuth
V11043	Extreme maximum wind direction	azimuth
V13011	Average precipitation	0.1mm
V13012	Daily precipitation >=0.1mm days	1day
V14032	Sunshine time	0.1h

From the above analysis, it was found that there is strong correlation between the average and maximum or minimum value of the climate-related variables, such as between the average temperature and average minimum temperature. For factor pairs with strong correlation, this study preferred to select the average value as input. The main reason is that extreme values only represent climate condition over a short period, while the average value is more representative of the climate condition during a certain period of time, namely, one month in our research.

3.3. Model Selection and Criterion

The objective of this study was to predict real-world fuel consumption rate of vehicles according to the vehicle factors and climate condition. The selection of the model's input factors has already been described in the above section. The proposed models are introduced in this part.

Because fuel consumption rate predicted by the model is a continuous variable, we choose regression modelling technique in this paper. The regression models used in this study are the linear regression, Naïve Bayes regression, Neural Network regression, Decision Tree regression, and LightGBM models.

The criterion for model selection include the mean absolute error ($\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}|y_i - y_i'|$), mean absolute percentage error ($\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\frac{|y_i - y_i'|}{y_i}$), mean squared error ($\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - y_i')^2$), and R squared ($\mathrm{R2} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$).

In the above formula, y_{i} denotes the true value, $y_{\mathrm{i}}^{'}$ denotes the predicted value, and $\overline{y_{\mathrm{i}}}$ denotes the mean actual value. Smaller MAE, MAPE, and MSE, and larger R2 mean that the error between the predicted and the actual value is smaller, which indicates that the model fits well and performs better.

4. Results

4.1. Model Training and Experiment Results

After removing the missing values, outliers, and standardization from the original data, 70% of the data were selected as the training dataset and the remaining 30% of the data were used as the test dataset. The training and testing results are presented in Table 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

Table 5. Results of regression model training and testing.

Model	Training data				Testing data			
	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2
refConsumption	1.897	26.4%	5.365	-2.244	1.894	26.4%	5.352	-2.243
Linear regression	0.993	11.3%	1.763	0.593	0.991	11.4%	1.767	0.596
Naïve bayes	1.029	11.8%	1.853	0.573	1.027	11.8%	1.851	0.577
Neural network	0.988	11.8%	1.722	0.603	1.004	12.1%	1.796	0.590
Decision tree	1.052	12.0%	1.929	0.555	1.047	12.0%	1.916	0.562
lightGBM	0.861	9.8%	1.354	0.690	0.911	10.4%	1.536	0.642

In Table 5, the 'refConsumption' row represents the result from directly using MIIT reference fuel consumption rate as model prediction. As can be seen, the error between the reference fuel consumption rate value and real-world fuel consumption rate value is quite large. The remaining rows represent the training and prediction errors of the four regression models, respectively.

4.2. Comparison and Analysis of Different Models

In this section, we compare the five different models by our proposed criterion.

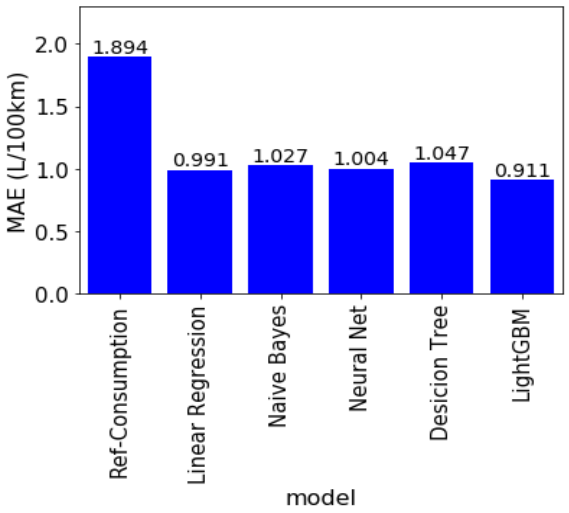

Figure 1. MAE of different models.

Figure 1 shows the mean absolute error (MAE) between the model prediction and actual values of each model. As can be seen, the mean absolute error from the reference fuel consumption rate provided by MIIT is 1.894 L/100 km, while the mean absolute error by our dataset including vehicle factors and climate condition is approximately 1 L/100 km. Among them, the mean absolute error of LightGBM model (0.911 L/100 km) is the lowest.

However, the MAE only indicates the absolute value of deviation and cannot reveal the magnitude of relative deviation from actual values. Therefore, Figure 2 shows the mean absolute percentage error (MAPE) between the model prediction and actual values. The results reveal that the MAPE between the reference fuel consumption rate and the real-world fuel consumption rate is approximately 26.4%. The best prediction model is still LightGBM and the corresponding MAPE is 10.4%, which is higher by 16% compared with the reference rate. This demonstrates that our proposed prediction model could be applied practically in the prediction and revision of vehicle fuel consumption.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

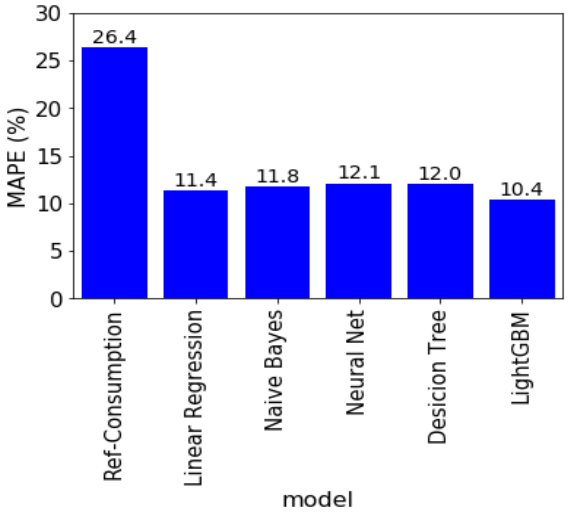

Figure 2. MAPE of different models.

Additionally, the mean square errors (MSE) of the model prediction and actual value of different models are shown in Figure 3. The results reveal that the mean square error of the LightGBM model is the lowest. This verifies our proposition that LightGBM performs best among the five regression models.

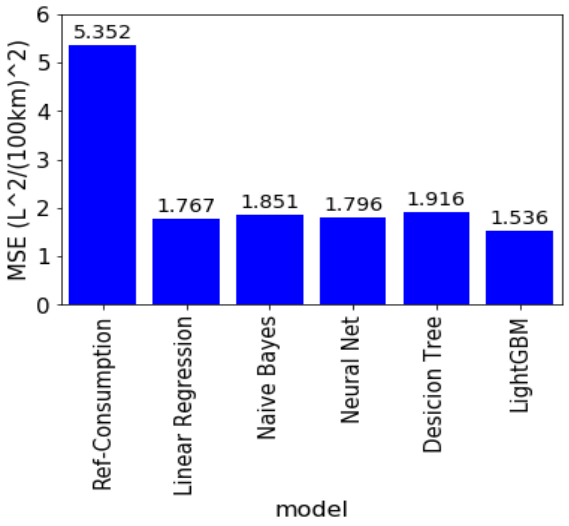

Figure 3. MSE of different models.

Notably, R2 is an index measuring the degree to which an independent variable explains a dependent variable in a regression model. From Figure 4, it can be seen that the reference fuel consumption rate by MIIT explains real-world fuel consumption rate to a very low degree, while the highest R2 value is that of the LightGBM model, followed by that of the linear regression model. The result indicates that dependent variables in LightGBM model explain 64.2% of the variation in independent variable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

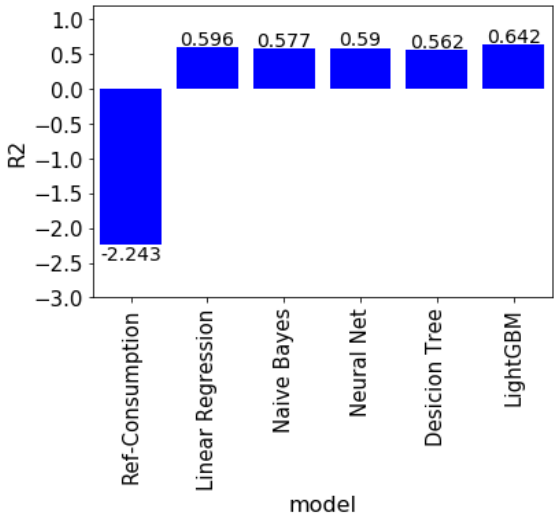

Figure 4. R2 of different models.

From the above results, it can be seen that the prediction error of real-world fuel consumption rate by the five regression models based on the vehicle and climate parameters are much lower than that of the reference fuel consumption rate value provided by MIIT. The MAPE can be reduced by 16% by most, which shows that the proposed prediction model has practical significance and can be applied in real-world applications.

Moreover, the comparison between the results obtained by the different models reveals that LightGBM regression model is the optimal one with best performance in reducing prediction error.

5. Discussion

The above results reveal that the LightGBM model achieved the best performance. The estimated weights of the input parameters in the LightGBM model are shown in Figure 5.

From the relative weights of the different factors, it can be seen that the reference fuel consumption rate is the highest, which is consistent with reality that MIIT reference value could present a large part of the actual fuel consumption.

Second, the engine power and vehicle brand are input factors with weights exceeding 0.1, which indicates that vehicle parameters and brand are the main impact factors for real-world fuel consumption.

Additionally, the engine displacement and average pressure (V10004) are input factors with weights exceeding 0.05, which indicates that the importance of air pressure, which may be attributed to the great effect of atmospheric pressure on the combustion efficiency of gasoline fuel.

Moreover, among the climate factors, the average temperature (V12001), average wind speed (V11002), and sunshine times (V14032) also exert great impact on real-world fuel consumption rate.

In summary, this part carried out comparative analysis of the vehicle and climate factors in our dataset and found that, in addition to the reference fuel consumption rate, the vehicle factors that have greater impact on the real-world fuel consumption rate are the vehicle brand, engine power, and engine displacement. The climate factors that have greater influence on real-world fuel consumption rate are the average air pressure, average temperature, and sunshine time.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2021 doi:10.20944/preprints202111.0029.v1

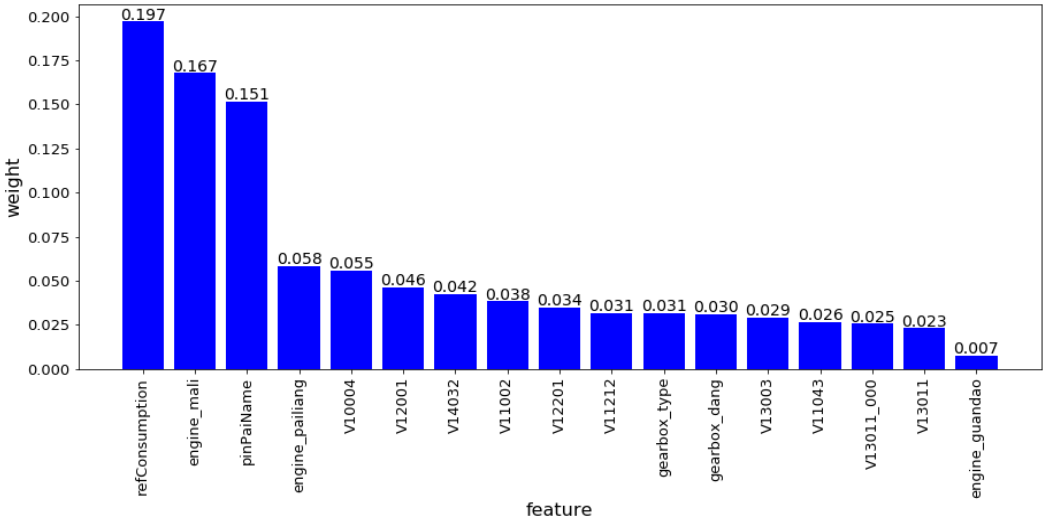

Figure 5. Weights of 17 Factors in LighGBM Model.

6. Conclusions

With the ongoing innovation and development in information technology, artificial intelligence (AI) will greatly accelerate technological progress in our increasingly digital and data-driven world. In this paper, we utilize five regression models, namely, the Linear Regression, Naïve Bayes regression, Neural Network regression, Decision Tree regression, and LightGBM models, to estimate real-world fuel consumption rate of light-duty vehicles in China, based on a large set of individual real-world driving and fuel consumption data.

The MAE, MAPE, MSE, and R2 between real-world fuel consumption rate and the value predicted using the vehicle and climate factors were far better than barely referring to the fuel consumption rate provided by MIIT of China. The comparison of the different models reveals that LightGBM regression model performs best among the candidate models according to all our criterion (MAE=0.911 L/100 km, MAPE=10.4%, MSE=1.536, R2=0.642).

This study also assesses 17 different factors, and determines the priority ranking of each factor. From the relative weight of each factor LightGBM model, it can be seen that the three most important factors are reference fuel consumption rate, engine power, and vehicle brand.

Author Contributions: Conceptualization, Isabella.Zeng, J.Xiong and T.Wu.; methodology, Isabella.Zeng and S.Tan; formal analysis, Isabella.Zeng and S.Tan; resources, Y.Li and T.Wu; data curation, T.Wu; writing—original draft preparation, Isabella.Zeng; writing—review and editing, X.Ding, J.Xiong; funding acquisition, Y.Li and T.Wu. All authors have read and agreed to the published version of the manuscript.

Funding: The project was sponsored in part by the National Natural Science Foundation of China (71804181, 61902037), in part by the Fundamental Research Funds for the Central Universities under Grant 500419804, and in part by the National Center for Mathematics and Interdisciplinary Sciences, CAS.

Acknowledgments: We thank Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript.

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

References

1. Peng, T., Ou, X.; Yuan, Z.; Yan, X., et al. Development and application of China provincial road transport energy demand and GHG emissions analysis model, *Appl Energy* **2018**, *222*, 313-328,
2. Liang, X.; Zhang, S.; Wu, Y.; Xing, J., et al. Air quality and health benefits from fleet electrification in China. *Nat Sustain* **2019**, *2(10)*, 962-971.
3. Jiang, S.; Chen, H. Examining patters of scientific knowledge diffusion based on knowledge cyber infrastructure: A multi-dimensional network approach. *Scientometrics* **2019**, *121*, 1599-1617.
4. Li, Y.; Tang, G.; Du, J.; Zhou, N.; Zhao, Y.; Wu, T. Multilayer perceptron method to estimate real-world fuel consumption rate of light duty vehicles. *IEEE Access* **2019**, *7*, 63395-63402.
5. Li, X.; Zhang, Q.; Peng, Z.; Wang, A.; Wang, W. A data-driven two-level clustering model for driving pattern analysis of electric vehicles and a case study. *J Clean Prod* **2019**, *206*, 827-837.
6. Wu, T.; Han, X.; Zheng, M. M.; Ou, X.; Zhang, X. Impact factors of the real-world fuel consumption rate of light duty vehicles in china. *Energy* **2019**, *190*, 116388.
7. Kwang-Ho, Ko.; Moon, Byung-Koo.; Lee, Soo-Hyung.; Tong-Won.; Won-Ho, et al. An economic calibration method for fuel consumption model in hdm4 (vol 89, pg 959, 2016). *Wirel Pers Commun* **2016**, *89*, 959-975.
8. Jiao, X.; Bienvenu, M. Field measurement and calibration of HDM-4 fuel consumption model on interstate highway in Florida. *Int J Transp Sci Technol* **2015**, *4*, 29-46.
9. Perrotta, F.; Parry, T.; Neves, L.C.; Buckland, T.; Benbow, E.; Mesgarpour, M. Verification of the HDM-4 fuel consumption model using a big data approach: A UK case study. *Transp Res D Transp Environ* **2019**, *67*, 109-118.
10. Yamashita, R.J.; Yao, H.H.; Hung, S.W.; Hackman, A. Accessing and constructing driving data to develop fuel consumption forecast model. In IOP Conference Series: Earth and Environmental Science, Proceedings of the 3rd International Conference on Advances in Energy Resources and Environment Engineering, Harbin, China, 8-10 December 2017; IOP Publishing, 2018; p. 012217.
11. Ahn, K.; Rakha, H.; Trani, A.; Van Aerde, M. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *J Transp Eng* **2002**, *128*, 182-190.
12. Rakha, H.; Medina, A.; Sin, H.; Dion, F.; Van Aerde, M.; Jenq, J. Traffic signal coordination across jurisdictional boundaries: Field evaluation of efficiency, energy, environmental, and safety impacts. *Transp Res Rec* **2000**, *1727*, 42-51.
13. Lei, W.; Chen, H.; Lu, L. Microscopic emission and fuel consumption modeling for light-duty vehicles using portable emission measurement system data. *World Acad Sci Eng Technol* **2010**, *66*, 918-925.
14. Wang, H.; Fu, L.; Zhou, Y.; Li, H. Modelling of the fuel consumption for passenger cars regarding driving characteristics. *Transp Res D Transp Environ* **2008**, *13*, 479-482.
15. Ma, H.; Xie, H.; Huang, D.; Xiong, S. Effects of driving style on the fuel consumption of city buses under different road conditions and vehicle masses. *Transp Res D Transp Environ* ***2015***, *41*, 205-216.
16. Ejsmont, J.A.; Ronowski, G.; Świeczko-Żurek, B.; Sommer, S. Road texture influence on tyre rolling resistance. *Road Mater Pavement Des* **2017**, *18*, 181-198.
17. Kono, T.; Fushiki, T.; Asada, K.; Nakano, K. Fuel consumption analysis and prediction model for "eco" route search. In Proceedings of the 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting ITS America ERTICOITS Japan Trans Core, New York, NY, USA, 16-20 November 2008.
18. Hu, J.; Wu, Y.; Wang, Z.; Li, Z.; Zhou, Y.; Wang, H.; Bao, X.; Hao, J. Real-world fuel efficiency and exhaust emissions of light-duty diesel vehicles and their correlation with road conditions. *J Environ Sci* **2012**, *24*, 865-874.
19. Brundell-Freij, K.; Ericsson, E. Influence of street characteristics, driver category and car performance on urban driving patterns. *Transp Res D Transp Environ* **2005**, *10*, 213-229.
20. Wang, S.; Zhang, X.; Cao, J.; He, L.; Stenneth, L.; Yu, P.S.; Li, Z.; Huang, Z. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. *ACM Trans Inf Syst* **2017**, *35*, 1-40.
21. Zahabi, S.A.H.; Miranda-Moreno, L.; Barla, P.; Vincent, B. Fuel economy of hybrid-electric versus conventional gasoline vehicles in real-world conditions: A case study of cold cities in Quebec, Canada. *Transp Res D Transp Environ* **2014**, *32*, 184-192.
22. Weilenmann, M.; Favez, J.Y.; Alvarez, R. Cold-start emissions of modern passenger cars at different low ambient temperatures and their evolution over vehicle legislation categories. *Atmos Environ* **2009**, *43*, 2419-2429.
23. Alvarez, R.; Weilenmann, M. Effect of low ambient temperature on fuel consumption and pollutant and CO2 emissions of hybrid electric vehicles in real-world conditions. *Fuel* **2012**, *97*,119-124.
24. Xu, B.; Chen, X.; Li, K.; Hu, M.; Bian, Y.; Yu, Q.; Wang, J. Double-layer speed optimization for reducing fuel consumption with vehicle-to-infrastructure communication. *J Intell Transp Syst* **2019**, *23*, 1-12.
25. Wang, Q.; Huo, H.; He, K.; Yao, Z.; Zhang, Q. Characterization of vehicle driving patterns and development of driving cycles in Chinese cities. *Transp Res D Transp Environ* **2008**, *13*, 289-297.
    Gong, Q.; Midlam-Mohler, S.; Marano, V.; Rizzoni, G. An iterative markov chain approach for generating vehicle driving cycles. *SAE Int J Engines* **2011**, *4*, 1035-1045.