Estimating the Energy Consumption and Driving Range of Electric Vehicles with Machine Learning

Yong Wang¹,², Wei Chenlong³, He Hongwen ¹,², *
¹National Engineering Laboratory for Electric Vehicles, Beijing Institute of Technology, Beijing 100081, China
²Collaborative Innovation Center of Electric Vehicles in Beijing, Beijing Institute of Technology, Beijing, 100081, China
³School of Mechanical Engineering, Anhui Polytechnic University, Wuhu 241000, China

The first author’s e-mail:17862709675@163.com,
The corresponding author's e-mail: hwhebit@bit.edu.cn

Abstract. The Data-driven methods have been widely used in the SOC, SOH and energy estimation of electric vehicles, and they are recognized as the most promising approaches. However, the popular machine learning methods used for electric vehicles are often “black boxes” which result in the poor interpretation of the model. In this paper, a highly efficient gradient boosting decision tree (LGBM) is proposed to accurately estimate the driving range of electric vehicle. In this model, the feature importance scores are provided to discover the relationship. In addition, the proposed LGBM-T model is capable of generalizing the abstractions by taking into account the key features of time and temperature. Experimental results illustrate that the proposed LGBM-T algorithm is able to reproduce the driving mileage trajectory, with a low mean absolute error (MAE) bounded by 1.681% on real-world vehicles under complex operating conditions. The availability of this algorithm is further corroborated by comparing to the support vector machines (SVM) based estimator.

1. Introduction
The goals of net-zero emissions and carbon neutrality before 2060 were adopted by more than 100 countries. In recent years, the transportation sector has been the biggest source of greenhouse gas emissions in some developed countries. Therefore, Electric vehicles (EVs) play a pivotal role in reducing carbon pollution and having captured much attention by governments as a national strategy to develop EV. However, range anxiety is the critical issues faced of EVs that have not gained significant popularity among the general public. On the one hand, the limited charging infrastructure and the limited driving range restrict daily use. On the other hand, the estimation of the remaining driving range is often inaccurate and causing the inevitably shorter driving range, and it is pertain to battery energy management (BMS) of electric vehicles which is one of the key technologies.

If the remaining driving range cannot be accurately predicted, the drivers experiencing range anxiety fear that the state of charge of battery will run out before reaching the destination. Therefore, accurately predicting the remaining driving range and energy consumption estimation for EVs is very
important and it is a basic parameter in the BMS of EVs. Consequentially, several works were recently
targeted at improving the estimation of the driving range.

One way is to design a more accurate range estimation algorithm, so a variety of estimation
approaches have been developed over the years that can be roughly classified into two categories:

1. Model-based estimation methods: The main principle of these methods is related to the
mathematical models including vehicle model, battery model and road model with the energy
consumption model, it will output energy consumption rate parameter by establishing a nonlinear
estimation algorithm for remaining driving range. For example, if the energy consumption rate
parameter is calculated by the nonlinear estimation algorithms which is 11 Kilowatt-hours per hundred
kilometers, and the total battery capacity of an EV is 40 Kilowatt-hours, it concluded that the electric
vehicle can reach up to 440 kilometers of range. However, the mathematical models were usually
simplified with many assumptions, and range and efficiency aren't directly related, which could cause
inaccurate prediction results.

2. Data-driven estimation methods: Data-driven methods only use the state variables measurements
as the input and mileage as the output to map a model between inputs-output. Learning the knowledge
from the raw data automatically, data-driven method is cost-effectiveness, commonality and reliability
in comparison with the model-based estimation methods. In [1], the authors propose a deep
convolutional neural network model for estimation the energy consumption considering three external
parameters. At the same time, the recent advances in machine learning have spurred the development of
data-driven battery SOC estimation [2], charging energy prediction [3], battery SOH estimation [4].

In addition, the research about the affecting parameters of electric vehicle energy consumption is
considered. Xie et al. [5] studied the effect of environmental parameters on energy consumption of
electric vehicles. The work presented in [6], the influencing factors of energy consumption, including
vehicle, environment, and driver-related factors, are studied based on the real-world driving data.

Due to limitations of algorithmic capabilities, the existing methods applied to driving range of
electric vehicles estimation/prediction are generally overlooking the black-box model explainability. To
address this issue, in this paper we propose an ensemble machine learning methodology. More
specifically, we incorporate the state-dependency characteristics of a state-of-the-art battery into the
range estimator and provide an overall methodology to build a model and extract accurate range
predictions in a realistic scenario. In this study, we collected the real-world data operating in four
different months, and introduced the LightGBM regression method to establish a model to estimate the
remaining driving range.

This paper is structured as follows: In Section 2, the data preparation step is explained. Section 3
presents the driving range prediction framework and the prediction results. Finally, Section 4 is
dedicated to the conclusion.

2. Data

2.1. Data description
At present, the data under real-world driving conditions has become a key performance index for electric
vehicle (EV) drivers. The abundance of available real word data can contribute to the further theory
research of electric vehicles driving range estimation. In this work, the real-world vehicular operation
data are collected from the National Monitoring and Management Platform for New Energy Vehicles,
which is formed by the national monitoring center for new energy vehicles and auto manufacturers in
China. The big data platform for EVs can provide round the-clock monitoring service for more than 3
million EVs by February 2020.

During the data collection process, it collected by vehicle terminal contains the aspects of vehicle
driving information and battery system states, which produced by different electric taxis with the same
model. The dataset consists of data of one vehicle that powered by the lithium iron phosphate battery
(LiFePO4) battery pack and the data sampling interval for recording of vehicles is 10s. The vehicle is
operated in Beijing and both driving state and charging state are included. The vehicle data collector for
track and record the state of vehicle operation, real-time data time, instantaneous speed, etc. The data of the battery system includes: maximum and minimum cell temperatures, total voltage and current, motor voltage and current of controller, and SOC.

The next step needs to select training and validation data for the data-driven models. To simulate the complex real-world EV battery loading behaviors, 80% of EV driving profiles are used to battery modeling and data training. The training data is collected from different months, including January, April, July and November 2018 - literally 300,000 data points, they are arranged in time order in rows. As are listed in Table 1, The samples of training data points track the date time (Time), speed, the total voltage of the vehicle (V), The total current of the vehicle (C), the state of charge (SOC), maximum cell temperature (T_max), minimum cell temperature (T_min), the motor voltage (M_v), the motor current (M_c) and total mileage every 10 s. The validation data was produced by different vehicles with the same model of training data, using the universal data format.

Table 1  RECORD SAMPLE OF THE RAW DATA POINTS

| Time          | Speed | V      | C | SOC | T_max | T_min | M_v | M_c | Mileage |
|---------------|-------|--------|---|-----|-------|-------|-----|-----|---------|
| 20180104180816 | 1     | 527.9  | 15| 50  | 20    | 14    | 527 | 7   | 34307   |
| 20180104180826 | 23    | 521.6  | 108| 50  | 20    | 14    | 521 | 98  | 34307   |
| 20180104180836 | 36    | 523.5  | 47 | 49  | 20    | 14    | 526 | -11  | 34307.1 |
| 20180104180846 | 8     | 530.3  | -20| 49  | 20    | 14    | 529 | -20 | 34307.2 |
| 20180104180856 | 9     | 525.7  | 53 | 49  | 21    | 15    | 523 | 83  | 34307.2 |
| 20180104180906 | 40    | 522    | 67 | 49  | 21    | 15    | 523 | 28  | 34307.3 |

2.2. Data processing

For data-driven model, accurate training data will produce more accurate prediction results, that is, data quality affects the model performances. In addition, the acquisition of real data is usually encountered by difficulties and challenges in contrast to standard drive cycles, because the original raw dataset may contain abnormal data points. And the raw data from the big data platform are characterized by high volume, low density, error. Therefore, data processing is a necessary process before training the model. The abnormal points in the raw data includes missing data, duplicate records, different formats, mistakes in data measurement. As the sporadically occurring device measurement device limitation and data error.

To ensure the quality of the data, the data processing is introduced as the process of correcting anomalies in the raw data. By analyzing the datum of dataset, there are three main types of abnormal data: missing value, duplicate records and outlier data. The first step removes all duplicate records and the second step is to detect and to remove outliers. The third step is processing the missing value records in the dataset. It is often disadvantageous to get rid of the observations that have missing data, in most cases, the missing data also has valuable information. Therefore, we need to infer those missing values from the existing part of the data and our processing methods are mean imputation.

3. Methodology and Results

3.1. LightGBM regression

LightGBM [7] is an ensemble machine learning algorithm based on the Gradient Boosting Decision Tree developed by Microsoft team in 2017. It has been extensively used in data modelling competitions like Kaggle because of its highly efficient operation and perfectly accurate performance. The principle of LightGBM regression (LGBM) is to build a strong learner with hundreds of decision tree learners. That said, the LGBM is a decision tree ensemble model that consists of a number of classification and regression trees. By using the error residual of previous weak learners as the next learning, a new regression tree is generated. The sum of predicted value from the multiple decision trees would be the
final predicted output, and a typical flow chart of LGBM to estimate the remaining driving range (DR) of Electric Vehicles is illustrated in Fig. 1.

The data set \( \mathcal{D} = \{(x_i, y_i)|[\mathcal{D}]| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\} \) containing \( n \) data points and \( m \) features in the training data, \( X \) is assumed as \( \{x_i\} \). \( DR_i \) represents the true value of driving range on i-th sample point. Then, the estimation \( DR_i^* \) of the ensemble model with k-trees can be obtained by:

\[
DR_i^* = \sum_{k=1}^{K} f_k(x_i) \tag{1}
\]

where \( f_k \) denotes an independent tree structure, and \( f_k(x_i) \) represents the score of the i sample point by the k-th tree. In order to train the regression tree model by learning the functions \( f_k \), we define the objective function:

\[
J = \sum_{i=1}^{n} l(DR_i, DR_i^*) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}
\]

where \( \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \) is the regularization term. It controls the complexity of the model that tend to select both a simple and predictive model. To avoid over-fitting, the objective function needs to trade-off between accuracy and complexity by the loss function and regular term. Here \( l \) is a training loss function, \( \gamma \) and \( \lambda \) are penalty coefficient for adjusting the number of leaves \( T \) and the leaf weights \( w \), respectively. By adding one new tree at time step \( t \) as \( f_t(x) \), the prediction of the \( t \) round iteration is trained:

\[
DR_i^*(t) = DR_i^*(t-1) + f_t(x) \tag{3}
\]

After the objective function is transformed into k round iteration by greedy algorithm, we take the Taylor expansion of the loss function up to the second-order, equation (2) can be expressed as follows:

\[
J^{(t)} = \sum_{i=1}^{n} l(DR_i, DR_i^*(t-1) + f_t(x_i)) + \Omega(f_t)
\]

\[
= \sum_{i=1}^{n} [g_t f_t(x_i) + \frac{1}{2} h_t f_t^2(x_i)] + \Omega(f_t) \tag{4}
\]

where, the first-order differential \( g_t \) and second-order differential \( h_t \) of the loss function are defined as:

\[
g_t = \frac{\partial}{\partial DR^*(t-1)} l(DR_i, DR^*(t-1))
\]

\[
h_t = \frac{\partial^2}{\partial DR_t^*(t-1)} l(DR_i, DR^*(t-1)) \tag{5}
\]

Figure 1 Illustration of LGBM based driving range prediction framework.

3.2. Estimation results
The MSE (Mean Squared Error), MAE (Mean Absolute Error), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the R2 (R-Square) are used to evaluate the driving range
estimation performance of the proposed framework. The smaller MSE, RSME, MAE and MAPE values indicate higher modelling accuracy, the larger R2 values indicate higher modelling accuracy.

To evaluate the model accuracy and prediction effect of the trained LGBM model, the prediction results of driving range shown in Fig. 2(a), which illustrates that the mileage errors are within a very small range, and predicted DR values are almost coincident with the actual values. And we could see the most prediction inaccuracy is around 700 time-steps, because the vehicle is stopped running during this time. The diagram of feature importance scores is shown in Figure 2(b), the feature importance scores denote the number of times that a feature is used for splitting in the training process. A higher feature importance score indicates that the corresponding feature is more important. Therefore, the SOC is valuable and should be highlighted. By comparison, the mouth is not a necessary feature in the model and may be abandoned to speed up the training process of the model.

![Figure 2](a)  Estimated driving range and the feature importance by proposed LGBM architectures.

In order to evaluate the effectiveness of the proposal, a test between the LGBM model and other popular machine learning models is further carried out. Firstly, the performance of LGBM model is compared with the support vector machines (SVM). Summary of the driving range prediction errors using different methods are illustrated in Table. 2, from which we can see LGBM has the highest prediction accuracy, even higher than the SVM-based prediction model. Therefore, LGBM-based method is very beneficial for us to implement driving range prediction for electric vehicle.

The LGBM-F model indicates using feature engineering to the LGBM models, which is the process of transforming raw data into features that better represent the DR estimation. Table 2 reveals also that, as expected, the time feature of “day-of-year” and “temperature” have positive and statistically significant effect on DR. Accordingly, the state of healthy of battery and the driving environment is relevant to the energy consumption rate and driving range. It can be found that, the LGBM-F method has performed significantly better than other models by the indexes MSE, RMSE, MAE, R2 and MAPE, because the LGBM-F based driving range estimation takes into account external information.

| Method   | MSE(%) | RMSE(%) | MAE(%) | R2   | MAPE(%) |
|----------|--------|---------|--------|------|---------|
| SVM      | 6.038  | 2.457   | 2.166  | 0.978| 3.288   |
| LGBM     | 5.014  | 2.358   | 2.081  | 0.973| 2.548   |
| LGBM-F   | 3.751  | 1.937   | 1.681  | 0.986| 2.309   |

4. Conclusions
In this paper, based on the real-world driving data that combined with environmental and battery information of electric vehicle in Beijing, an ensemble machine learning algorithm based on the Gradient Boosting Decision Tree for predicting the drivable range and period for vehicles under energy
remaining is presented. Firstly, the raw data are cleaned, and processed, it is ready for analysis. Meanwhile, the important features of electric vehicle operating states have been extracted. With the extracted key influencing factors, the relationship between the features and driving range has been explained by the feature importance scores. At last, the LGBM-F algorithm is proposed to interpret the prediction results of driving range, and it indicates that the proposed model has a good estimation accuracy. The presented method could predict accurately in drivable range and drivable period estimations. This leads to a better utilization of the battery energy and increases the effective driving range. From the practical experiment, we also find that the data driven based estimation method demand more advanced data preprocessing technologies and expert knowledge into the creation of feature extractors for robust against overfitting and accurate analysis.

Acknowledgements
The authors gratefully acknowledge the financial support from National Key R&D Program of China (Grand No.2017YFB0103700).

References
[1] Modi S, Bhattacharya J, Basak P. Estimation of energy consumption of electric vehicles using Deep Convolutional Neural Network to reduce driver’s range anxiety. ISA transactions, 2020, 98: 454-470.
[2] Tian J, Xiong R, Shen W, et al. State-of-charge estimation of LiFePO4 batteries in electric vehicles: A deep-learning enabled approach[J]. Applied Energy, 2021, 291(3):116812.
[3] Yang, Z et al. Data-driven framework for large-scale prediction of charging energy in electric vehicles. Applied Energy, 2021, 282: 116175.
[4] Lei C, J Meng, Stroe D I, et al. Multi-objective Optimization of Data-driven Model for Lithium-ion Battery SOH estimation with Short-term Feature[J]. IEEE Transactions on Power Electronics, 2020, PP(99):1-1.
[5] Xie Y, Li Y, Zhao Z, et al. Microsimulation of electric vehicle energy consumption and driving range[J]. Applied Energy, 2020, 267.
[6] Wang H, Zhang X, Ouyang M. Energy consumption of electric vehicles based on real-world driving patterns: A case study of Beijing[J]. Applied Energy, 2015:710-719.
[7] KE, Guolin, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, 30: 3146-3154.