Battery Health Prediction of New Energy Vehicles Based on LightGBM

Haolong Li

Smart City College of Beijing Union University, Beijing Union University, Beijing, 100101, China
2023240388018@buu.edu.cn

Abstract. In the era of rapid development of new energy vehicles, the health status of lithium-ion batteries (SOH) will have a direct impact on the endurance and safety of new energy vehicles, so it is very important to predict the health status of lithium-ion batteries. This study uses lightgbm as a data prediction model to study the prediction of battery health status of new energy vehicles. The influence factors of lithium-ion battery health status are taken as independent variables, and the performance of random forest, xgboost and lightgbm algorithms is systematically compared. The final prediction results show that the lightgbm model has the highest degree of fitting, and the R-squared value is as high as 0.9933, and the operation efficiency is also significantly higher than the other two models. The research shows that lightgbm can be used as an efficient solution for the SOH prediction of new energy vehicle batteries, which provides a strong scientific basis for the health management and maintenance of new energy vehicle batteries, and also provides an important reference for the optimization of battery management system, so that the new energy vehicle industry can develop better.

Keywords: Lightgbm; Prediction; Battery;health status.

1 Introduction

Driven by the global energy transformation and the goal of reducing carbon emissions, in order to solve the problems of the energy crisis and environmental pollution, new energy vehicles have developed rapidly. According to statistics, the sales and ownership of new energy vehicles in China have increased year by year [1]. As the core power source of new energy vehicles, lithium-ion batteries are widely used in new energy vehicles with their advantages of high energy density, long cycle life, light weight and high charging efficiency. The performance and health status of lithium-ion batteries directly affect the vehicle's endurance, safety and economy. However, the battery will undergo a complex electrochemical aging process in the long-term use process, resulting in capacity attenuation, internal resistance increase and other problems. When the battery capacity decays to 70%-80% of the target capacity, the performance of the battery will be seriously affected, which may cause the battery to fail to work normally [1]. Accurate prediction of battery health status is very important

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_10

to optimize battery management, extend service life and reduce maintenance costs. It can also reduce the probability of traffic accidents caused by battery failure and ensure the safety of users' lives and property. Thus, predicting the battery health status of new energy vehicles holds great significance.

Currently, the approaches for predicting battery health status can be primarily categorized into model-based methods, data-driven methods, and fusion model-based methods [2]. The model-based method is mainly through the mathematical model simulation of the battery aging mechanism, but its model complexity is high and requires a lot of manpower and material resources. The method based on the model is mainly to collect, process and analyze the battery data, mine the law from the historical data, and use the traditional model to predict. The method based on model fusion is to combine several models to improve the prediction accuracy. For instance, Kant A and colleagues put forward a combination of bidirectional recurrent neural networks and long short-term memory networks to boost the ability to predict the health state of lithium-ion batteries, which comes with complex pattern recognition capabilities and higher prediction accuracy [3]. Cai and Liu combined transform techniques with Long Short-Term Memory networks, discerning crucial characteristics by analyzing voltage, energy, and temperature curves across both time and frequency domains.By leveraging the transformer's inherent self-attention capabilities and the LSTM's proficiency in capturing long-term dependencies, they developed an effective fusion framework aimed at enhancing the precision and reliability of SOH forecasting [4].

Traditional prediction methods and model fusion methods have some shortcomings and limitations. As an efficient gradient lifting decision tree framework, lightgbm plays an important role in similar research problems with its advantages of high speed, low memory consumption and high accuracy. Xin Guomao and others used the lightgbm algorithm to process large-scale data sets and high-dimensional features such as voltage, current and power factor of street lamps, which accelerated the training speed of the model, found street lamp faults in time, greatly reduced the cost of maintenance and improved the efficiency of maintenance [5]. Niu et al. Proposed the lightgbm probability model, which is used for multi-step prediction of the number of sunspots in advance, reducing the prediction error and enabling effective anomaly detection during abnormal events [6]. Biswas et al. Studied and used eight machine learning algorithms to evaluate the accuracy of power load forecasting. The results show that the lightgbm model is the most accurate, and the improvement of accuracy has significant economic benefits for Bangladesh, which helps to minimize energy loss and optimize power system operation [7].

Therefore, this study proposes the application of the lightgbm algorithm in the prediction of battery health state of new energy vehicles, aiming to build a prediction model with high accuracy and strong robustness. Through the extraction of key features in the process of battery aging, combined with the integrated learning characteristics of lightgbm algorithm, accurate prediction of battery health status is achieved, which provides a scientific basis for the battery health management and maintenance strategy of new energy vehicles, and then promotes the healthy and sustainable development of the new energy vehicle industry.

2 Dataset and Method

2.1 Data Source and Description

The data set in this paper is the battery health data set of electric vehicles, considering the known key factors affecting battery performance. There are 10000 records in this data set, including 8 variables. The data volume is sufficient and complete to support subsequent analysis. This dataset is an ideal choice for machine learning regression tasks, especially to predict the percentage of battery health. It can also be used for exploratory data analysis, the construction of a battery health monitoring prototype, and research related to electric vehicle data modeling.

2.2 Index Selection and Description

Table 1 summarizes the basic information of each feature. In this study, seven factors affecting the battery health status are selected, which are the total vehicle distance, the number of completed battery charging cycles, the average speed, the average battery temperature, the percentage of total charging completed by rapid charging, the outdoor ambient temperature, and the average driving time. The battery health status is used as the dependent variable to predict. The range of each index is reasonable, there is no obvious abnormal value, and the distribution is nearly symmetrical, which can be used as an effective feature basis for battery health state prediction.

Table 1. Feature information table.

Feature name	minimum value	Maximum	average value	standard deviation	median
total_distance_km	1000.5	39988.9	20272.2	11217.6	20208.6
average_trip_speed_kmph	20.0	60.0	40.2	11.6	40.2
ambient_temperature_C	10.0	45.0	27.5	10.0	27.5
trip_duration_min	10.0	89.9	49.9	23.1	50.0
charging_cycles	50.0	787.0	369.5	203.5	366.0
fast_charging_ratio_%	0.2	89.6	28.7	15.9	26.6
average_battery_temperature_C	25.0	50.0	35.4	8.8	34.5
battery_health_%	45.0	100.0	68.4	16.5	71.2

2.3 Method Introduction

lightgbm algorithm is an improvement and upgrade of xgboost. The underlying mathematical principle is similar to xgboost. The algorithm process is shown in Figure 1.

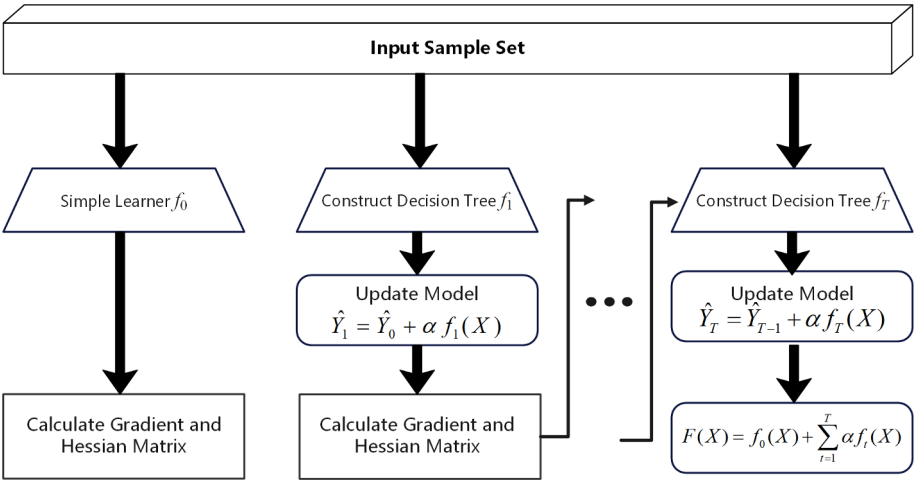

Fig.1. lightgbm algorithm flowchart (Picture credit: Original).

But many new methods are proposed based on xgboost. In terms of data preprocessing, the lightgbm modeling process will carry out data compression in three aspects. First, continuous variables will be discretized into boxes on the whole sample, and then the discrete features and discrete continuous variables will be brought in at the same time. The mutually exclusive feature binding algorithm will be used to reduce the dimension. This method can well overcome the problem of a large amount of information loss caused by the traditional dimension reduction method, and finally the gradient based unilateral sampling method will be used for down sampling.

In terms of decision tree optimization modeling, leaf wise is a more efficient strategy, as shown in Figure 2. Find the leaf with the largest splitting gain from all the current leaves each time.

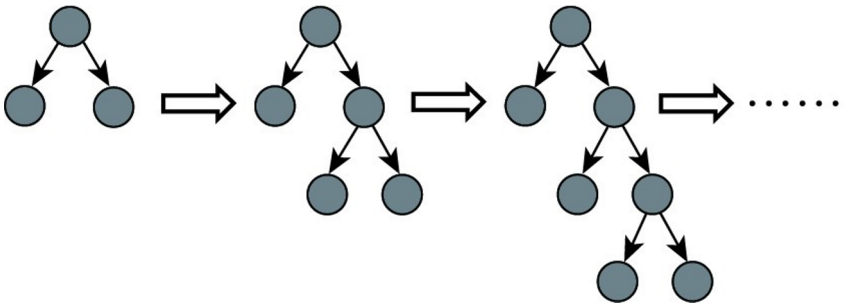

Fig.2. Based on the leaf splitting diagram [8].

The histogram in Figure 3 is used to accelerate the operation. By discretizing the continuous eigenvalues into histograms, the amount of calculation and memory consumption are greatly reduced.

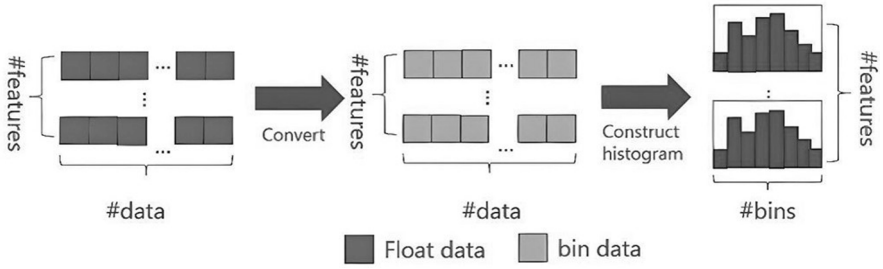

Fig. 3. Schematic diagram of histogram algorithm [9].

Because a large amount of data will be generated in the study of battery health status, this study uses a lightgbm model, which has the advantages of fast training speed, low memory occupation, efficient processing of large-scale data, and guaranteed accuracy to effectively predict battery health status.

3 Results and Discussion

3.1 Feature Importance Analysis

Figure 4 Feature weight analysis chart is a method to reveal the importance of different features by quantifying their contribution to the prediction results of the model. Its core role is to enhance the interpretability of the model, optimize feature selection and improve the performance of the model. It can be seen from the feature weight diagram that the weight of the total charge percentage completed by the fast charging method is the highest, which has the most significant impact on the model output, and is the key influencing factor for the prediction of battery health. The weight of average travel time is low, indicating that its contribution to battery health state prediction is weak.

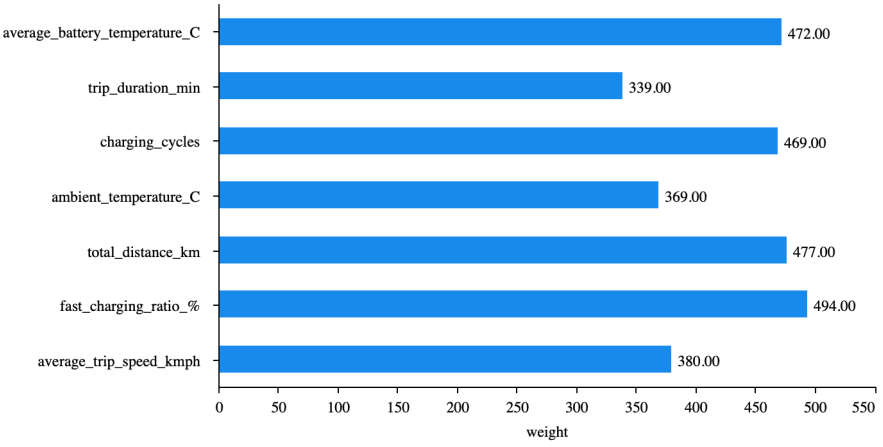

Fig. 4. Feature weight graph (Picture credit: Original).

Figure 5 provides an intuitive and unified feature importance analysis by quantifying the marginal contribution of each feature to the model prediction. The SHAP summary chart reveals the following key information through the contribution direction and intensity of features to the model output: from the vertical direction, charging_cycles and total_distance_km are at the top, indicating that they have the most significant global impact on the prediction results. Average_trip_speed_kmph effect is relatively weak. Moreover, the number of battery cycles, the total distance traveled by the vehicle, the average speed and the proportion of charging completed in a fast way are negatively correlated with the battery health state. The greater their values, the lower the battery's health state value.

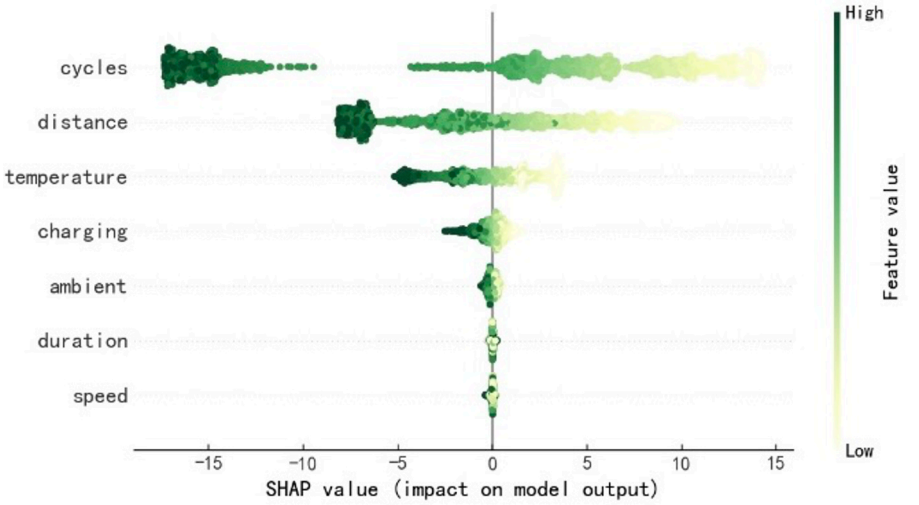

Fig. 5. SHAP outline (Picture credit: Original)

3.2 Model Evaluation

Figure 6 shows the linear relationship between the predicted value and the real value of the model. It can be seen that the linear distribution of data points indicates that the predicted value and the real value are closely related, and the R^2 value reaches 0.99, indicating that the model has a high degree of fitting, good fitting effect, and strong ability to predict the battery health state.

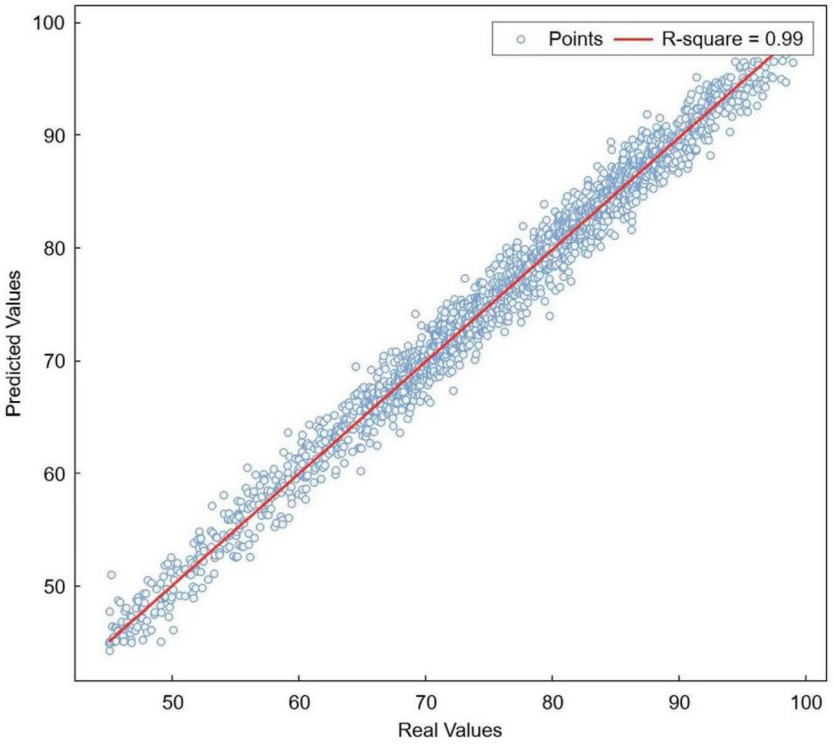

Fig. 6. Fitting regression diagram (Picture credit: Original)

The abscissa of Figure 7 is the test sample point, and the ordinate is the numerical value. The blue dot represents the real value, and the red asterisk represents the predicted value. It can be seen from the figure that most of the predicted values are close to the real value distribution, indicating that the model has a good prediction effect, but there are also some point deviations, reflecting that the model has errors in the prediction of these samples.

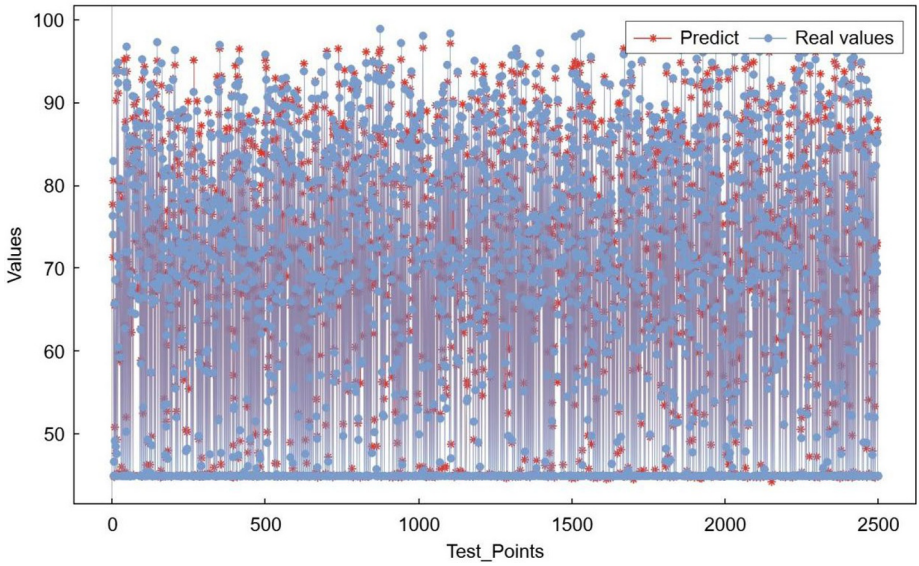

Fig. 7. Comparison between predicted value and real value (Picture credit: Original)

3.3 Comparison with Random Forest and Xgboost

The advantages of random forest are simple training, high parallel efficiency, strong anti over fitting ability, and good robustness to missing values and outliers, but the disadvantages are high model complexity, large memory consumption, and generally lower prediction accuracy than the boosting algorithm [10]. xgboost achieves higher prediction accuracy through gradient lifting and regularization, and supports user-defined loss function and feature importance analysis, but it has high computational overhead, slow training speed, and is sensitive to parameter tuning. lightgbm is further optimized on the basis of xgboost. It uses a histogram algorithm and unilateral gradient sampling to significantly improve the training speed and reduce memory consumption. It is especially suitable for large-scale data and high-dimensional features. At the same time, it supports direct processing of category features. However, its leaf wise growth strategy may increase the risk of over fitting, and its performance may not be as stable as xgboost on small data sets. In the prediction of the battery health state of new energy vehicles, lightgbm can balance efficiency and accuracy, showing excellent performance.

In order to highlight the advantages of lightgbm and explain the reasons for choosing the lightgbm model, this study uses the same data set to conduct prediction experiments with the other two models at the same time. Through the comparison in Table 2, it can be seen that the average absolute error value and R square value of lightgbm are slightly higher than the other two models, but the mean square error is lower than the other two models, indicating that the overall prediction accuracy of lightgbm in the current study is equivalent to that of the comparison model, and it performs better in controlling the square error and explaining the change law of target variables, with good prediction

stability and fitting effect. And lightgbm has a significantly higher processing speed than the other two models in the processing and prediction of large-scale data without losing the fitting effect and accuracy. The lightgbm algorithm can play a better role in the prediction of the battery health state of new energy vehicles and show better results.

Table 2. Model evaluation comparison table.

Model name	MAE	MSE	R^2
Random forest	0.9694	1.9232	0.9931
xgboost	0.9682	1.8873	0.9932
lightgbm	0.9705	1.8709	0.9933

4 Conclusion

In this study, a new energy vehicle battery health state prediction model is built based on the lightgbm algorithm. The comparative experiment verifies its superior performance compared with random forest and xgboost. For high-dimensional data such as batteries, lightgbm's leaf wise growth strategy can more effectively capture the nonlinear characteristics and local mutation points in the process of battery performance degradation. The experimental results show that under the same data set conditions, lightgbm can ensure the prediction accuracy and the speed of its training model is much faster than the other two algorithms, which improves the prediction accuracy and efficiency of battery health state of new energy vehicles. Due to the large number of samples in this study, the lightgbm algorithm can reflect obvious advantages, but in the small sample analysis, the accuracy of the xgboost and random forest algorithms may not be worse than that of lightgbm, so it is recommended to adopt a differentiation strategy for different scenarios in practical applications. Future research can increase the number of features and samples, and further explore the improved lightgbm algorithm combined with an attention mechanism to improve the detection sensitivity of battery abnormal state, or integrate multiple models to further improve the prediction accuracy and efficiency by using the advantages of different models.

References

1. Wang, Y., Ni, Y., Zheng, Y., et al.: Remaining Useful Life Prediction of Lithium-Ion Batteries Based on ALO-SVR. Proceedings of the Chinese Society for Electrical Engineering 41(4), 1445–1457, 1550 (2021)
2. Wang, N., Liu, X., Chen, Z.: A Review on Lithium-Ion Battery Lifetime Prediction. Electrical Appliance and Energy Efficiency Management Technology, 11, 1–13 (2018)
3. Kant, A., Kumar, M., Sihag, S.: Prediction of Electric Vehicle Battery State of Health Estimation Using a Hybrid Deep Learning Mechanism. International Journal of Green Energy 22(10), 2065–2078 (2025)
4. Cai, X., Liu, T.: State of Health Prediction for Lithium-Ion Batteries Using Transformer–LSTM Fusion Model. Applied Sciences 15(7), 3747 (2025)

5. Xin, G., Gu, X., Zhu, Y., et al.: Lighting Fault Prediction Method Based on LightGBM. Smart City 1–7 (2025)
6. Niu, B., Huang, Z.: Multi-Step Probabilistic Forecasting for Sunspot Numbers Based on LightGBM. Advances in Space Research 75(11), 8398–8410 (2025)
7. Biswas, S. S., Kundu, A., Chakrabartty, J.: LightGBM-Driven One-Day-Ahead Electrical Load Forecasting for Enhancing Operational Efficiency of a Power Distribution Company of Bangladesh. Electrical Engineering, prepublish, 1–24 (2025)
8. Feng, Q., Zhang, J., Zhu, J., et al.: Prediction of Stratum Fracture Width Based on LightGBM Algorithm. Energy and Environment Protection 47(01), 65–72 (2025)
9. Cui, H., Zhang, P., Zeng, X., et al.: Flight Landing Time Prediction Based on LightGBM. Automation and Instrumentation 02, 33–36 (2025)
10. Simatupang, C. J., Susanty, M.: Comparative Performance Analysis of Random Forest and Multilayer Perceptron Algorithms for Earthquake Magnitude Prediction in Indonesia. IOP Conference Series: Earth and Environmental Science 1521(1), 012021 (2025).