XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction

Yuyi Zhang
Faculty of Applied Mathematics and Control Processes
Saint-Petersburg State University
Saint Petersburg, Russia
Lesliezhang0825@gmail.com

Feiran Xu
Faculty of Applied Mathematics and Control Processes
Saint-Petersburg State University
Saint Petersburg, Russia
xfr97120@gmail.com

Jingying Zou
Saint-Petersburg State University;
General Development Department
Huawei Technologies Co. Ltd.
Saint Petersburg, Russia
jesco4482@gmail.com

Ovanes L. Petrosian
Saint-Petersburg Electrotechnical University "LETI";
Saint-Petersburg State University;
General Development Department
Huawei Technologies Co. Ltd.
Saint Petersburg, Russia
petrosian.ovanes@yandex.ru

Kirill V. Krinkin
Saint-Petersburg Electrotechnical University "LETI"
Saint Petersburg, Russia

2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT) | 978-1-6654-4534-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/NEURONT53022.2021.9472817

***Abstract*—The results of evaluating explanations of the black-box model for prediction are presented. The XAI evaluation is realized through the different principles and characteristics between black-box model explanations and XAI labels. In the field of high-dimensional prediction, the black-box model represented by neural network and ensemble models can predict complex data sets more accurately than traditional linear regression and white-box models such as the decision tree model. However, an unexplainable characteristic not only hinders developers from debugging but also causes users mistrust. In the XAI field dedicated to "opening" the black box model, effective evaluation methods are still being developed. Within the established XAI evaluation framework (MDMC) in this paper, explanation methods for the prediction can be effectively tested, and the identified explanation method with relatively higher quality can improve the accuracy, transparency, and reliability of prediction.**

Keywords—XAI evaluation, black-box model explanations, neural network, ensemble models

I. INTRODUCTION

The XAI evaluation framework has been paid more and more attention under the condition that the XAI methods have been widely accepted. At present, the interpretability of the model is obviously focused on the field of ML and DL, especially when their models are used in medical, financial, military and other fields that require high transparency and trust, the interpretability of the model is even necessary. However, black-box models such as neural networks [1] and ensemble models [2] are not inherently interpretable. Therefore, the XAI method is absolutely needed to promote the more efficient application of black-box models with higher accuracy in real life. The XAI evaluation framework can select XAI methods that are more suitable for the model, which further improves the credibility of the black-box model.

The black-box models used for prediction mainly include neural network models and ensemble models. Among them, the ANN [3] model is used as a representative of the neural network model, LightGBM [4] and random forest [5] are used as the representative of the ensemble model, LightGBM is a model based on the boosting algorithm in the ensemble model, and the random forest is based on the bagging algorithm. Both boosting and bagging [6] are composed of many simple tree models. If there is no connection between the tree models, and the final output result through "voting" is the bagging algorithm. If there is a strong connection between these tree models, a boosting algorithm is formed, that is, the output of the previous tree model will affect the next tree model. In addition, linear regression and white-box model-decision trees are also used in prediction tasks for comparison.

SHAP [7] and LIME [8] are used as the method to explain the model. On the one hand, both belong to the feature attribution explanation. The same working logic ensures that they are comparable and the validity of the evaluation framework we propose. On the other hand, considering the wide application of them in tabular data, they are representative. At the same time, random disturbances are used as the baseline for comparison, thus objectively showing the effectiveness of the XAI method. In order to establish the final XAI evaluation framework, in this paper, the main work includes:

- Comparison of accuracy measures of black box models (ANN, LightGBM and random forest) with linear regression and decision trees;
- Use XAI method to explain the black-box model;
- Create a new XAI evaluation framework, and demonstrate the principles of the framework through visualization technology, and finally apply the framework in the form of an evaluation matrix to select the most suitable XAI method.

978-1-6654-4534-4/21/$31.00 ©2021 IEEE

II. Black-box Model Predictions and Explanations

A. Dataset Description

The data set used for prediction – Boston housing from "sklearn", the original data comes from the U.S. Census Bureau. The data set includes 506 rows of instances, 13 features, and 1 predicted variable. The data set information can be viewed at the link:

https://www.kaggle.com/jamieleech/boston-housing-dataset

B. Black-box Model Predictions

Neural network and ensemble models are typical representatives of black-box model. Among them, artificial neural networks (ANN) [3] and ensemble models such as Random Forest (RF) [5] and LightGBM [4] have excellent performance for prediction.

ANN is an adaptive nonlinear dynamic system composed of a large number of simple basic units (neurons) interconnected by activation functions. The structure and function of each neuron are relatively simple, however, system behaviors produced by the combination of neurons are extremely complicated and difficult to explain. ANN is divided into three layers: input, hidden, and output. The ANN architecture we built is shown in Table 1.

TABLE I. ANN Architecture

Layer (type)	*Output shape*	*Activation*
Dense	(None, 128)	ReLU
Dense	(None, 64)	ReLU
Dense (output)	(None, 1)	-

GBDT[9] is an extremely important model in the field of machine learning. Its main idea is to use weak classifiers (decision trees) to iteratively train to obtain the optimal model. This model has the advantages of excellent training effect and avoiding overfitting. LightGBM [4] (Light Gradient Boosting Machine) is a framework that implements the GBDT algorithm, which has the advantages of faster training speed, lower memory consumption, and better accuracy.

In a random forest, multiple decision trees are built, and they are merged together to obtain more accurate and stable predictions. The random forest classifier uses all the decision tree classifiers and the hyperparameters of the bagging classifier to control the overall structure. It should be noted that regression also has a random forest regressor.

In order to be able to visually see the advantages of the black-box model in prediction, the traditional linear regression (LR) and the classic white-box prediction model – decision tree (DT) are used as a control group to predict the Boston housing data set together with ANN, LightGBM and RF. The quality statistics of the prediction results are shown in Table 2.

TABLE II. The Quality Statistics

Model	**Metrics**		
	R^2	*MSE*	*MAE*
LR	0.64856	28.40585	3.69136
DT	0.74361	20.72322	3.05065
ANN	**0.79477**	**16.58769**	**2.57977**
LGBM	**0.80417**	**15.82852**	**2.53292**
RF	**0.81888**	**14.63918**	**2.35044**

Metrics shows that ANN, LightGBM and RF have almost absolute advantages in prediction tasks. However, compared with linear regression and decision trees, the prediction process of ANN, LightGBM and RF is incomprehensible by humans due to their internal complex structure. As a result, black-box models with higher accuracy cannot completely replace traditional models and white-box models with lower accuracy in many fields. Therefore, the development of effective methods to explain the black-box model is crucial.

C. The Meaning and Methods of XAI

Before starting to use the XAI method, the target task of model interpretation needs to be clarified, that is, the realization of the target of model interpretability. The interpretability of the model is by no means the full presentation of the internal working process of the black-box model, which is almost impossible to achieve and meaningless. The meaning of model interpretability is to allow the model to be safely applied to the corresponding field. In other words, model interpretability is also a tool for business guidance and case interpretation. Therefore, the meaning of model interpretability can be defined as:

- Business guidance: the identification of the most important features in the model to guide the business.
- Explain the case: Explain which feature is good or bad for a specific case.

SHapley Additive exPlanation (SHAP) can interpret complex machine learning models. Although it comes from game theory, it only uses this idea as a carrier.

- SHapley: Represents calculating the Shapley Value of each feature variable in each sample.
- Additive: Represents that the shapley value corresponding to the feature variable is additive for each sample.
- exPlanation: Represents the interpretation of a single sample, that is, how each feature variable affects the predicted value of the model.

The main idea of LIME is to use interpretability models (such as linear models, decision trees) to locally approximate the prediction of the target black box model. This method does not go deep into the model. By slightly perturbing the input, it detects what happens to the output of the black box model. According to this change, an interpretability model is trained at the original input point.

D. XAI: Explanation forBlack-box Model Predictions

In order to clearly show the XAI method's explanation of the black-box model prediction, take SHAP as an example to explain the prediction results locally and globally. SHAP's global explanation of LightGBM prediction is shown in Fig. 1. The 13 features are arranged in order according to the degree of influence on the prediction result. Taking the Boston housing data set as an example, according to the results of the global explanation, the government can more clearly understand the reasons that affect real house prices, which is helpful for policy macro-control.

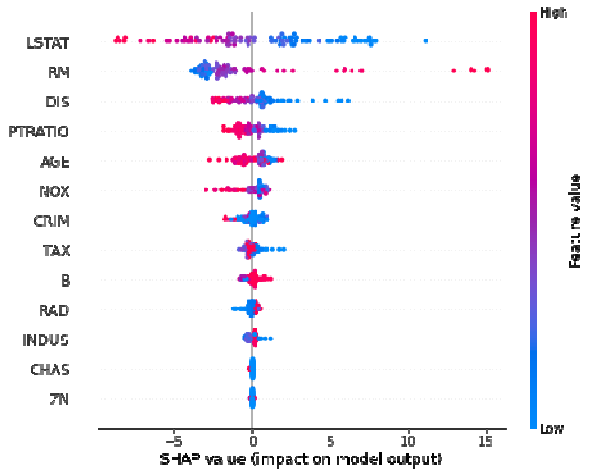

Fig. 1. SHAP global explanation of LightGBM prediction

SHAP local explanation of LightGBM prediction is shown in Fig. 2. Local explanation focuses on each instance, and outputs features that have an impact on each instance. According to local explanations, buyers can more accurately measure the target house price and make reasonable decisions.

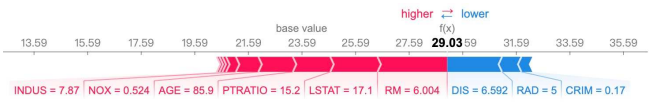

Fig. 2. SHAP local explanation of LightGBM prediction

In the actual application of XAI, different XAI methods will output different explanations. On the one hand, this situation is due to the different principles of various XAIs. On the other hand, because XAI explanation of the predictive model is specific, in the case of the same data set but different black-box models, different explanation results will be output. In order to solve this problem, the evaluation of the XAI method is necessary for the correct explanation of the predictive model.

III. XAI Evaluation Framework for Black-box Model Explanations

The establishment of the XAI evaluation framework is based on a premise: deleting or changing the large contribution (calculated by the XAI method) features in the data set will result in a significant decrease in the accuracy of the model's prediction. Therefore, based on this premise, the degree of change in the metrics (R2, MSE, MAE) can be used as the core of the evaluation framework.

A. XAI Evaluation Framework: Mean Degree of Metrics Change (MDMC)

In Perturb the original data according to the output of the XAI method, and input the modified data into the prediction model to obtain a new metrics (R2*, MSE*, MAE*), then the degree of change of the metrics (D) can be is defined as:

$$D = f(M - M^*) \tag{1}$$

M is the original metric, and M* is the changed metric.

Combining the degree of change of all metrics, the final evaluation framework (MDMC) can be defined as:

$$MDMC = \frac{1}{n}\sum_{i=1}^{n} D = \frac{1}{n}\sum_{i=1}^{n} f(M - M^*) \tag{2}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[(R2_0 - R2_i) + (MSE_i - MSE_0) + (MAE_i - MAE_0)\right]$$

It should be noted that the values of MSE and MAE in Equation (2) need to be used after normalization. In theory, the larger the value of MDMC, it means that the black-box prediction model has made significant changes to the data set, thus proving the effectiveness of XAI.

B. XAI Evaluation Framework: Visualization

In order to understand the MDMC framework intuitively, take the ANN model as an example to evaluate the XAI methods. The changes in the metrics are shown in Fig. 3. It is worth noting that if the stability of the ANN model itself is lacking, then randomly deleting features will also cause a decrease in the accuracy of the model's prediction. In order to eliminate this possibility, in the evaluation process, the data set of "randomly deleted features" was used as a comparison of the XAI methods to prove the stability of the ANN model we constructed.

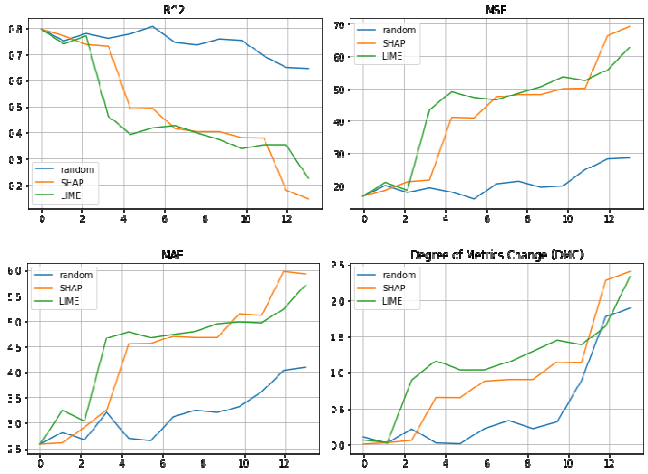

Fig. 3. Visualization of the changes in the metrics

As shown in Fig. 3, compared to randomly deleting features, the perturbation based on the results of the XAI methods significantly reduces the accuracy of the ANN prediction model.

From the metrics of view, when interpreting the ANN model, the performance of SHAP is more stable, and the trend of change is consistent; on the other hand, the performance of LIME is relatively unstable, however, from the early stage, the performance of LIME is significantly better than that of SHAP. These results mean that the LIME method is more sensitive to high-contribution features. Moreover, according to the sum of the degree of change of the three metrics (DMC), it is shown that when the three metrics are combined, LIME performs better overall.

C. XAI Evaluation Framework: Evaluation Matrix

Visualization can intuitively evaluate the XAI method, but for those tasks with high-precision requirements, a quantitative evaluation matrix is necessary. The process of the XAI methods on LightGBM and Random Forest is similar to that of ANN and follows the same calculation formula. The values in Table 3 are calculated based on Equation (2). The larger the value, the more effective the XAI method is.

TABLE III. EVALUATION MATRIX

MDCD	ANN	LightGBM	RandomForest
Random	0.8333	0.8222	0.6242
SHAP	1.4154	**1.5004**	1.3286
LIME	**1.6506**	1.3995	**1.6069**

The experimental results show that in the black-box model, the perturbation based on the XAI methods is more effective than the random perturbation. On the other hand, within the XAI method, LIME performs better in the ANN model and the random forest model, and SHAP is more effective for the LightGBM model.

IV. CONCLUSION

After comparing with the accuracy of traditional linear regression and white-box models, black box models such as neural networks and ensemble models including boosting algorithms and bagging algorithms have absolute advantages in prediction, however, the unexplainable nature makes the black-box model an obstacle in the process of practical application. The XAI methods can obviously alleviate this problem. By identifying features with relatively high contributions, users can understand the black box model more clearly when using the black box model to predict, thereby increasing trust. However, because many XAI methods have different principles and characteristics, different XAI methods output different results for the same black-box model. Therefore, the establishment of an evaluation framework for XAI methods is necessary. The XAI methods are evaluated through the established XAI evaluation framework – MDMC. The results show that LIME is more suitable for ANN model and random forest model based on bagging algorithm, and SHAP is more suitable for LightGBM based on boosting algorithm.

REFERENCES

[1] Aggarwal C.C. Neural networks and deep learning. Springer, 2018, 10: 978-3.

[2] Xiao Z, Wang Y, Fu K, et al. Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. ISPRS International Journal of Geo-Information, 2017, 6(2): 57.

[3] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research Journal of pharmaceutical and biomedical analysis, 2000, 22(5): 717-727.

[4] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, 30: 3146-3154.

[5] Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

[6] Wen L, Hughes M. Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. Remote Sensing, 2020, 12(10): 1683.

[7] Lundberg S, Lee S I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.

[8] Ribeiro M T, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier[C]. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.

[9] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001: 1189-1232.