A modified LIME and its application to explain service supply chain forecasting

Haisheng Li[1], Wei Fan[1], Sheng Shi[1], Qiang Chou[1]

[1] Lenovo research, Beijing, China
lihs6@lenovo.com
fanwei2@lenovo.com
shisheng2@lenovo.com
chouqiang1@lenovo.com

Abstract. Recently, researchers are focusing more on the study of explainable artificial intelligence due to its usefulness on various scenarios that request trust, such as deciding if one should trust a prediction, choosing between models, improving an untrustworthy model and identifying why a model should be trusted. One main research issues is how to improve the interpretability, while preventing any deterioration of accuracy of the model. For this issues the model-agnostic explanation method is a kind of solution. In the paper we propose a modified LIME algorithm based on locally fitted by decision tree regression called tree-LIME which is a model-agnostic method. Further, we clarify the fidelity measure definition in regression explanation problem by using mean absolute error (MAE). The experiments on real service supply chain forecasting application show that (1) our proposed approach can improve the fidelity of the explainer which lead to a more accurate explanations for individual instances and (2) our approach gives a more intuitive and visualized tree expression for explanation. (3) The approach also works well when applied to service supply chain forecasting.

Keywords: Explainable artificial intelligence, Model-agnostic, Local explanation, Decision tree.

1 Introduction

Explainable artificial intelligence, the objective of which is to explain the black-box machine learning models, is an important topic in machine learning research and its applications. The interpretation of black-box model not only corresponds to assessing trust, but also relates to deploying a new model. Many sophisticated methods, especially deep neural networks and ensemble methods, are very complicated and even for human experts are struggle to interpret them. So interpretable machine learning has been a resurgence in recent years.

There are a number of methods to perform model explanation. In general these approaches can be categorized into two types: model-agnostic methods[1-5] and model-

specific methods[6-9]. (i) Considering the explanation problem, model-agnostic methods firstly learn an interpretable model that locally or globally[10] approximates the given model. Then, using the learned interpretable model to explain given model. (ii) Considering the explanation problem, model-specific method is to judiciously design representation form of the algorithm and make the algorithm explainable in itself. In fact, frequently used linear model and decision tree essentially belong to this types. Whether model-agnostic method or model-specific method have their advantages and drawbacks. The model-agnostic method faced the fidelity problem, e.g. the explainable model cannot effectively approximate original model. While for the model-specific method a tradeoff often must be made between accuracy and interpretability[11]. Besides, except these two typical category approaches there are still other explainable methods. In [12], depending on the natural language processing method, the author train a deep neural network with picture-text pair to explain the content of the picture. In [13], author interpret deep neural network GANs by identifying a group of interpretable units and visualize to interpret.

In this paper, we focus on the model-agnostic method and we think that for model-agnostic method the fidelity should be measured. If an explainable model cannot effectively approximate original model, then explanation is imprecise. Specifically, we explore a new kind of LIME (local interpretable model-agnostic explanation) based approach called tree-LIME to perform explanations. LIME is a local interpretable model-agnostic explanation method proposed in [14-16]. It explains the complicated machine learning model by locally approximate the model being explained in explainable feature space. The LIME is local linear and it can explain the predictions of any classifier and regressor.

In our approach, we modify the locally linear model of LIME to decision tree regression and by clarify the fidelity definition for regression we apply our approach to service supply chain forecasting which is a time series forecasting problem and is modeled to regression problem. So the main contributions are as follows.

- Tree-LIME, a modified method base on LIME, which can effectively locally approximate the original model to be explained with the tree interpretable representation.
- Clarify the fidelity definition for regression between explainable model and original model by computing mean absolute error (MAE).
- Applying the propose method to explain the service supply chain forecasting and show its advantage in fidelity and representation.

2 Tree-LIME and regression fidelity

2.1 Locally fitting by decision tree regression

Model-agnostic explainable method LIME is a method that locally approximate the original model in explainable feature space. In the local region of feature space, LIME use a linear model to explain the samples needed to be explained. In our method, we use decision tree regression to approximate and explain samples. Specifically, define

model g ∈ G as an explanation, where G is a class of potentially interpretable models. Let the model being explained be denoted f : $R^d \rightarrow R$. In regression, f(x) is the response variable. LIME produces explanation as following:

$$\xi(x) = \frac{argmin}{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (1)$$

where, $L(f, g, \pi_x)$ is fidelity component and π_x is a proximity measure between an instance to x which defined the local region to be fitted. $\Omega(g)$ is interpretability component.

For original LIME, the author select a sparse linear model K-LASSO as the explainable model g. In our method we use decision tree regression model CART as the g. The changes for the local explainable model will lead two kind of effects. Firstly using a nonlinear tree model replace liner model will increase local fidelity. Secondly instead of using a linear model as an interpretable representation, the replacement leads to tree formation representation. As we will show in experiment later, locally nonlinear fitting can improve the fidelity and tree representation for explanation is transparent and concise. Before that, we will define fidelity measure for regression.

2.2 MAE as the Fidelity measure

For the Model-agnostic explainable method, approximation to original model is important. The fidelity is to measure this kind of property. However for classification and regression, the fidelity will be varied. In classification it have been defined in several researches. While for regression, there is still no definition.

In [10], the author define the fidelity for classification as the percentage of test-set examples on which the classification made by an explainable model agrees with its original counterpart model. It can be formally defined as:

$$Fidelity_{classification} = \frac{N_{f=g}}{N} \quad (2)$$

Where, f represents original model and g is explainable model. N is the size of test dataset of the original model f and $N_{f=g}$ represents the number that explainable model agrees with original model on test dataset.

For regression problem, the mean absolute error (MAE) is a usually used evaluation metrics and in this paper we compute the MAE between the explainable model's result and original model's result as the fidelity measure. So it can be formally defined as:

$$Fidelity_{regression} = MAE_{g,f} = \frac{1}{n}\sum_{i=1}^{n}|g_i - f_i| \quad (3)$$

Where, g_i is the explainable model's approximate result and f_i is original model's forecasting result. In experiment chapter, we will use this definition to comparing two explainable model's fidelity for regression problem.

3 Experiment

In our experiment, we compare our proposed approach with original LIME on service supply chain forecasting data which is to forecast the usages of each week for computer repairing parts, such as mainboard, hard-drive and LCD panels. We purified this data from real-world application of service supply chain and preprocess this time

series data to the form of tabular data, so we can model this time series forecasting problem to a regression problem. After the extraction, the dataset contains 271242 train samples and 25068 test samples. Then, we train an ensemble model on this dataset which is the ensemble with two varied XGBoost models. For this ensemble it is hard to explain even for machine learning practitioners. We use the proposed tree-LIME and original LIME to explain this ensemble, then compare the Fidelity measure and show the interpretable representations.

3.1 About service supply chain data

In details, service supply chain forecasting data are for predicting the usage quantities of computer's repairing parts, which is prepared for customer's repairing. For large computer manufacturers the service supply chain usually maintain thousands of repairing parts for their customers. The supply chain's planner are tasked with predicting their weekly usages for up to one quarter (13 weeks) or half a year (26 weeks) in advance. The usage quantities of repairing parts are influenced by many factors, including the parts commodity, the product segmentation and the machine types. With hundreds of individual planners predicting usages based on their unique circumstances, accuracy of results can be quite varied. In our extracted dataset, historical usages for 5414 parts are provided, so we have history usages of 5415 parts to form train dataset. In test dataset we need to forecast the usages of 2136 parts for up to one quarter (13 weeks). Because in every week we can extract one sample for training and testing dataset, the train dataset contain 271242 instances and test dataset have 25068 instances to be forecasted. According to the problem, 10 attributes are extracted so far and details are as follows:

1. TopmostPN: parts number, e.g. ID
2. IB: install base which means the quantity of computers in warranty
3. Commodity: the commodity types of the parts
4. BU: the business unit that the part belongs to
5. Segment: the segmentation of product, which allocate the correspond parts
6. IB_duration: the IB's duration weeks from first the first part in warranty
7. Usage_duration: the duration weeks that first usage occurred
8. Year: which year that the usages occurred
9. Month: which month that the usages occurred
10. WeekofYear: which week of the year that the usages occurred

3.2 Fidelity on service supply chain forecasting

In this chapter, we compare proposed tree-LIME with the original LIME using Fidelity measures. We randomly pick up 50 instances out of 25068 instances from test dataset. Firstly we forecast these 50 instances by original XGBoost ensemble. Then we explain these instances separately by tree-LIME and LIME. At last, computing the Fidelity measure of tree-LIME and LIME based on XGBoost ensemble result. To conquer randomness, we repeat this experiment 3 times. In experiment, the parameter of tree depth

is set 4. The reason we do not do this experiment on whole 25068 test dataset is that the explanation process is time consuming. Table1 shows the fidelity values for three times experiments. It can be seen from Table1 that for all three time experiments on pickup samples, fidelity values of tree-LIME is lower than original LIME. We can come to a conclusion that the fidelity performance of tree-LIME is better than original LIME which it is benefitted from nonlinearly fitting in local feature space. In fact for a complicated dataset from real life application, the boundary of feature space are usually nonlinear.

Table 1. Comparition of Fidelity

Fidelity	Experiment-1	Experiment-2	Experiment-3
LIME	9.64	6.43	11.28
Tree-LIME	6.22	3.66	3.63

3.3 Analysis about the depth of the tree

In our approach, one of the important parameters is the depth of the tree. The tree depth is the important parameter because it can adjust the tradeoff of the explainer's fidelity and its interpretability. When the depth is too deep for a tree, it becomes hard to interpret. However, we think when the depth of the tree is shallow, tree model will degenerated to liner model and will reduce fidelity. In the following experiment, we validate that the depth of the tree will effect fidelity. We design two groups experiment. For each group we randomly picked 50 instances and set tree depth as 3, 4 and 5 separately to compute fidelity value. As shown in Table2 when the tree depth is 5, the fidelity value of both groups achieved the best fidelity performance. The experiments validate our conjecture about the effect of tree depth. In real application considering the user's limitation, some users may accept the tree of 4 layers while the others may accept that up to 10 layers. we recommend this parameter is set to 4 or 5.

Table 2. Comparition of the tree depth's effect

Fidelity	Depth=3	Depth=4	Depth=5
Group-1	7.33	6.22	4.89
Group-2	3.93	3.66	3.02

3.4 Interpretability and the tree representation

Considering interpret representation in various applications and for different users, a linear model[17], a decision tree, a decision rule list may or may not be interpretable. In the following we show the difference of the interpretable representation between our approach and original LIME. When we change the linear fitting method to the nonlinear tree fitting method in the local explainable space, the interpretable representation is changed correspondingly. We perform further study on representation cases between LIME and tree-LIME. Fig1 and Fig2 show the explanation result of an instance in service supply chain forecasting which is forecasted by our XGBoost ensemble either. For

this instance, the forecasting result of our ensemble is 4 pieces of usage, the approximation result of LIME is 0.236≈0 pieces and our tree-LIME is 4.824≈5 pieces. The parameter of tree depth for tree-LIME is set to 4. Again from the view of fidelity, tree-LIME's fidelity is 1, LIME is 4 and tree-LIME is better than LIME. In the following we concentrate on the explanation representations of two methods. The explanation representation result is shown in Fig1 and Fig2.

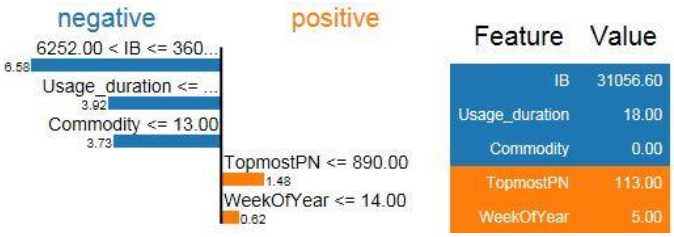

Fig.1. LIME representation

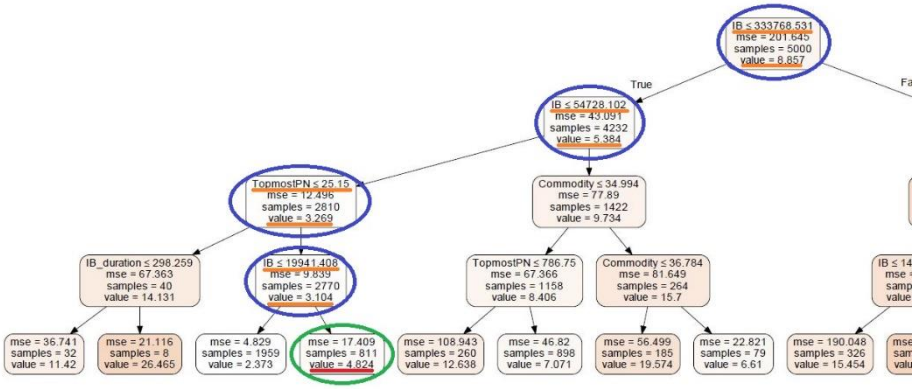

Fig.2. Tree-LIME representation

In Fig1, LIME uses the weight of the linear model as the interpretable representation output. The positive weight means the corresponding feature have the positive effect for the regression result and vice versa. As shown in Fig1, LIME's explanation result can be translated that if 6252.00<IB<=36063.00, Usage_duration<38.00, Commodity<=13.00, TopmostPN<=890.00 and WeekOfYear<=14.00, then forecasting result is 4 pieces. In Fig2, tree-LIME leads to a tree interpretable representation. Decision tree is inherently explainable and it is the decision rule in essence too. From Fig2 we can find that the forecasting result is 4 pieces because 19941.408<IB<=54728.102 and TopmostPN>25.15.

So the problem is which kind of representation is better? As shown in Fig1 and Fig2, The interpretable representations obtained by LIME and tree-LIME are indistinguishable, so it is hard to say which is better. For representation problem, there have not been adequate study in different representations and it is likely that different representations

are appropriate for different kinds of users and domains[18]. From the view of transparent and concise interpretation, the translated decision rules of tree-LIME is less than LIME, meanwhile tree-LIME is more fidelity than LIME. We think LIME and tree-LIME can both explain the service supply chain forecasting result well and the explanation of LIME and tree-LIME are both reasonable for real service supply chain planers. The advantage of our proposed approach is that the approach obtained more concisely explanations at the same time that the fidelity is higher.

The appropriation problem of explanation representation is really hard to evaluate. Considering decsion rule is a kind of general thinking mode, we think in practice the explanation should be sample, consice and easy to translate to decision rules.

One not mentioned problem is categorical features problem for regression. In Fig1 and Fig2, the 'Commodity' feature is an example. For regression the general way to cope with categorical feature is the integer encoding, but it is troublesome when explaining the model for tree representation in our approach. It is meaningless that a categorical feature is greater than some values or less than some values. For this problem, selecting other tree representation may be a good choice and this will be for future works.

4 Conclusion and future work

This paper investigates the fidelity and interpretability representation of local model-agnostic explainer LIME. Although LIME can explain any classifier and regressor, its fidelity and interpretability representation for regression can be improved. Our major contribution is a modified approach called tree-LIME, which uses the tree representation and increases the capability of local approximation, e.g. fidelity. Further, for regression problem we clarify the fidelity definition by using mean absolute error (MAE) between explainable model and original model. Our experiments on service supply chain application demonstrate that the proposed approach can increase fidelity and explain the forecasting result well. In the future, we will cope with how to better explain categorical features, experiment on multiple datasets and evaluate interpretability representation with human subjects.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765-4774.

[2] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:.03888*, 2018.

[3] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279-288: ACM.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Nothing else matters: model-agnostic explanations by identifying prediction invariance," in *30th Conference on Neural Information Processing Systems*, Barcelona, Spain, 2016.

[5] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge information systems*, vol. 41, no. 3, pp. 647-665, 2014.

[6] Y.-Y. Chang, F.-Y. Sun, Y.-H. Wu, and S.-D. Lin, "A Memory-Network Based Solution for Multivariate Time-Series Forecasting," *arXiv preprint arXiv:.02105*, 2018.

[7] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675-1684: ACM.

[8] F. Wang and C. Rudin, "Falling rule lists," in *Artificial Intelligence and Statistics*, 2015, pp. 1013-1022.

[9] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350-1371, 2015.

[10] M. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *Advances in neural information processing systems*, 1996, pp. 24-30.

[11] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721-1730: ACM.

[12] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*, 2016, pp. 3-19: Springer.

[13] D. Bau *et al.*, "Gan dissection: Visualizing and understanding generative adversarial networks," *arXiv preprint arXiv:.10597*, 2018.

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

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," in *2016 ICML Workshop on Human Interpretability in Machine Learning* New York, USA, 2016.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349-391, 2016.

[18] S. Singh, M. T. Ribeiro, and C. Guestrin, "Programs as black-box explanations," in *30th Conference on Neural Information Processing Systems*, Barcelona, Spain, 2016.