







Evaluating Fidelity of Explainable Methods for Predictive Process Analytics

Mythreyi Velmurugan^(✉) , Chun Ouyang , Catarina Moreira ,
and Renuka Sindhgatta 

Queensland University of Technology, Brisbane, Australia
{m.velmurugan,c.ouyang,catarina.pintomoreira,renuka.sr}@qut.edu.au

Abstract. Predictive process analytics focuses on predicting the future states of running instances of a business process. While advanced machine learning techniques have been used to increase the accuracy of predictions, the resulting predictive models lack transparency. Explainable machine learning methods can be used to interpret black-box models. However, it is unclear how fit for purpose these methods are in explaining process predictive models. In this paper, we aim to investigate the capabilities of two explainable methods, LIME and SHAP, in reproducing the decision-making processes of black-box process predictive models. We focus on *fidelity* metrics and propose a method to evaluate the faithfulness of LIME and SHAP when explaining process predictive models built on a Gradient Boosting Machine classifier. We conduct the evaluation using three real-life event logs and analyze the fidelity evaluation results to derive insights. The research contributes to evaluating the trustworthiness of explainable methods for predictive process analytics as a fundamental and key step towards human user-oriented evaluation.

Keywords: Predictive process analytics · Explainable AI · Evaluation metrics · Explanation fidelity

1 Introduction

Predictive process analytics focuses on applying predictive analytics to forecast future states of business process executions [10]. While advanced machine learning techniques have been used to increase accuracy of process predictions, the resulting predictive models become ‘black-box’ models. Methods and techniques have been proposed in machine learning to explain black-box models, forming a new research theme known as explainable AI (XAI) [3]. Several recent studies in predictive process analytics (e.g., [2, 9]) have attempted to apply existing XAI methods to interpret black-box process predictive models. However, there have been few studies on evaluating how well available XAI techniques interpret process predictions. One key measure of explanation fitness is *fidelity*, which aims to determine how faithful the explanation is to the black-box predictive model, i.e., how well the explanation method can mimic the black-box model [3].

In this paper, we identify and draw on fidelity evaluation studies in XAI, and propose a method for evaluating explanation fidelity for process predictions. We apply the proposed method to evaluate the performance of LIME and SHAP in interpreting process predictive models built on XGBoost, which has been shown to be most accurate in process outcome predictions [10], and analyse the evaluation results to derive insights. The research contributes to evaluating the trustworthiness of explainable methods for predictive process analytics as a fundamental and key step towards human user-oriented evaluation.

2 Background and Related Work

2.1 Explainable AI

While more complex algorithms often produce more accurate results, it is harder for a human to understand their internal workings, thus becoming a ‘black box’ and requiring interpretation [3]. Post-hoc interpretation refers to the interpretation of a predictive model, a data neighbourhood or a prediction created after the model has been generated, typically by an interpretation mechanism external to the predictive model [3]. Two popular black-box-model-agnostic, local explanation methods in literature are LIME and SHAP. LIME determines the importance of all features in an input by perturbing the dataset to create a surrogate linear model that captures the black-box model’s behaviour at a specific neighbourhood [8]. SHAP uses a game theoretic approach to assign a value, known as SHAP value, to each feature at the instance level, describing its contribution to the final output—the prediction [5].

2.2 Fidelity of Explanations

While post-hoc explanation methods can be used to interpret complex models, because the explanation method is distinct from the prediction method, it is possible that the explanation generated is not always faithful to the decision-making of the original black box. Hence, it is important to understand the fidelity of the explanation method. Two ways of measuring fidelity are defined in [6]: external and internal fidelity. *External fidelity* measures the similarity of decisions made by a surrogate model or interpretation of a black box and the black box itself, but this does not measure the similarity of their decision-making processes – defined as *internal fidelity* [6]. A common method of evaluating the fidelity of post-hoc approaches is to remove or change features identified by the interpreter and comparing the changes in prediction probability of the black box [1, 4], though this is typically applied to image or text data.

2.3 Problem Statement

Explainable predictive process analytics has emerged as a new research topic, and current studies have attempted to use existing explainable methods in XAI [2, 9].

However, it is as yet unclear how fit for purpose these methods are for explaining process predictions, given the relatively complex multidimensional structure of event log data. As such, it is important to understand how well explainable methods can mimic the decision-making of process prediction black boxes. However, methods of assessing internal explanation fidelity for tabular data like event logs remain unexplored. In literature, internal fidelity evaluation methods often apply ablation, in which the most influential features are removed from the input [1]. However, this is typically applied to text or image data, where the “removal” of features is relatively simple, and would not be appropriate for tabular data. This has motivated us to draw on existing work to *build a method to assess the fidelity of post-hoc methods used in explainable predictive process analytics*.

3 Fidelity Evaluation Method

We evaluate the *internal fidelity* of explanations as we are interested in the fidelity of the interpreter’s decision-making processes, not the decision (i.e., prediction) itself. An ablation approach to measuring internal fidelity will not hold for tabular data like event logs, particularly when using XGBoost which automatically imputes missing data. As such, a *perturbation strategy* was judged to be more effective. For the prediction of each instance, ten explanations were generated—to mitigate the effects of explanation instability—and the top 10% of features that were most common in the explanations were identified. For each feature, LIME presents the feature value or feature value distribution which affected the black box’s prediction. For example, an explanation including “ $1 < \text{Activity_A} < 3$ ” indicates that the occurrence of “Activity_A” more than once, but fewer than three times was influential. As SHAP presents only the feature’s influence on the end result, feature value distributions were generated based on the SHAP value for a specific feature, for a specific instance (i.e., what feature values would produce similar SHAP values in the entire test dataset). For example, if “Activity_A” has a SHAP value between 0.5 and 0.6 only when the feature value is between 1 and 3, this would be the distribution attributed to a SHAP value of 0.54. Using these distributions, for each instance:

1. A prediction using input vector x was generated, along with the prediction probability for the predicted class $Y(x)$
2. For each feature to be perturbed, a new, uniform distribution outside of the existing distribution was created to draw new feature values from
3. For each feature to be perturbed, a new value was randomly sampled from the new distribution to replace the original value for that feature, creating the perturbed feature vector x'
4. The prediction probability for the originally predicted class was determined for input x' resulting $Y(x')$, and the difference between $Y(x)$ and $Y(x')$ was computed

Each instance was perturbed ten times, and the differences in prediction probability were used to calculate the mean absolute percentage error (MAPE) of

the differences as the fidelity score for each instance (see Eq. 1). We chose to perturb the feature vectors, instead of the event logs used to derive the feature vectors, as the input for both the black box and the explanation methods were the derived features, and not the original event log.

It is important to note that the definition of the measure in Eq. 1 focuses on the (local) explanations at the process instance level. The overall evaluation of such a measure over the entire event log can be calculated as the average of the scores for all instances in the event log. Error functions have previously been applied to quantify internal fidelity, averaged out over the size of a dataset [1]. As such, MAPE is used to measure the fidelity of explanations, and we calculate the fidelity (\mathcal{F}) of an explanation for a single process instance in an event log as follows:

$$\mathcal{F} = \frac{\sum_1^{|X'|} \frac{|Y(x) - Y(x')|}{Y(x)}}{|X'|} \quad (1)$$

where:

- x = original feature vector for the process instance
- X' = Set of perturbations for x and $x' \in X'$
- $Y(x)$ = Prediction probability given input x
- $Y(x')$ = Prediction probability given input x'

Note that this measure is naturally bounded by the fact that prediction probabilities fall between 0 and 1.

4 Evaluation and Analysis

4.1 Design of Experiments

The prediction target is process instance outcomes, as outcome classification is one of the most common process prediction problems. Since XGBoost is generally the most accurate algorithm for outcome prediction [10], it was used to create the underlying black-box models. The XGBoost classifiers were trained on different data encoding and bucketing methods. The following combinations of bucketing and encoding methods were used:

- Aggregate encoding for dynamic attributes with prefix-length bucketing
- Index-based encoding for dynamic attributes with prefix-length bucketing
- Aggregate encoding for dynamic attributes with no bucketing

In the “no bucketing” method, all data is compiled as one bucket and a single classifier is trained on this bucket. When prefix-length bucketing is used, data is grouped (bucketed) based on shared prefix length (the number of activities that have already been completed in a process instance), and a classifier is trained for each bucket. For example, in a dataset with process traces ranging from prefix length of 1 to 40, forty classifiers will be trained. Aggregate encoding, as the name implies, aggregates the details of the entire case into a summary, while index-based encoding attempts to preserve the temporal details of the case as much as

possible. As such, combining aggregate encoding with single bucketing preserves the characteristics within the original event log the least, while combining index-based encoding with prefix-length bucketing preserves the most.

SHAP and LIME, two popular post-hoc interpretation methods, were chosen for evaluation due to their prevalence in explainable predictive analytics [2,9].

4.2 Datasets

We use three open-source, real-life event logs, all varying in the amount of cases recorded, types of attributes present and context (see Table 1). We follow the preprocessing, bucketing and encoding methods used in [10].

Table 1. A summary of statistics of three event log datasets

Event log		Production ^a	Sepsis cases ^b	BPIC2012 ^c
Description		A manufacturing process	Hospital event log showing sepsis cases	Loan application process
No. of cases (before encoding)		220	782	4,685
Proportion of positive cases		55.0%	16.0%	53.4%
Maximum prefix length		23	29	40
Prefix lengths used		1–20	1–25	1–25
Feature vector shape	Single bucket & aggregate encoding	162	274	134
	Prefix-length buckets & aggregate encoding	Min: 137 Max: 156	Min: 153 Max: 218	Min: 43 Max: 134
	Prefix-length buckets & index-based encoding	Min: 100 Max: 844	Min: 147 Max: 535	Min: 11 Max: 1654

^a <https://doi.org/10.4121/uuid:68726926-5ac5-4fab-b873-ee76ea412399>

^b <https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460>

^c <https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f>

The Production dataset has the fewest cases and a substantial number of static and dynamic attributes (though more dynamic than static), and around 55% of the cases were completed with a positive outcome. The Sepsis Cases dataset is highly unbalanced with only 16% positive cases. It also contains considerably more static than dynamic attributes, which will result in comparatively longer feature vector lengths when using aggregate encoding, but comparatively shorter feature vector lengths at higher prefix lengths when using index-based encoding. The training dataset was balanced through downsampling, but the testing data remains unbalanced. The BPIC2012 event log contains one (numerical) static attribute and a number of dynamic attributes, most of which are categorical, resulting in comparatively smaller feature vectors when using aggregate encoding, but the feature vector size will increase considerably as prefix length increases when using index-based encoding.

4.3 Results and Analysis

The source code implementing the proposed evaluation method in Sect. 3 and the design of experiments in Sect. 4.1 as well as the associated experiment results are available at: <https://git.io/JIYtH>.

Evaluation Results. Overall, fidelity scores were low-to-moderate for both explainable methods (see Table 2). SHAP is the better performer, though generally by small margins. Both explanation methods were generally least faithful when single bucketing and aggregate encoding are used, whereas the best combination of bucketing and encoding that produced better results varied across datasets.

Table 2. Overall fidelity results for each of the three datasets

		Production	Sepsis cases	BPIC 2012
Single bucket aggregate encoding	LIME	0.26	0.36	0.37
	SHAP	0.27	0.46	0.41
Prefix-length buckets aggregate encoding	LIME	0.47	0.37	0.38
	SHAP	0.51	0.49	0.42
Prefix-length buckets index-based encoding	LIME	0.36	0.51	0.32
	SHAP	0.51	0.56	0.4

LIME and SHAP are almost comparable when evaluating fidelity. In many cases, such as with SHAP for the BPIC2012 dataset (see Fig. 1), the faithfulness of explanations varies across instances in an almost uniform distribution. This suggests that only some explanations are faithful, but there appears to be no pattern or trend of faithfulness with regards to prefix length, encoding method, bucketing method or the initial prediction probability.

There is generally no link between prefix length and fidelity, except with BPIC2012 (most noticeable in LIME), where a higher prefix length generally results in a more faithful explanation. The large size of the BPIC2012 dataset (at least in comparison to the other two) has resulted in more reasonable black-box accuracy at the higher prefix lengths (see Fig. 2), which in turn appears to have ensured that LIME’s surrogate models better fit the data.

Analysis and Findings. At first glance, these low scores seem to suggest that LIME and SHAP cannot accurately mimic process prediction black boxes. However, a further analysis indicates that the reasons for poor fidelity results may lie in the following aspects.

Firstly, event logs are inherently complex due to their multidimensional nature and variety of event attributes, and it is possible that the processing of the event log to a simpler, algorithm-readable feature vector could have led

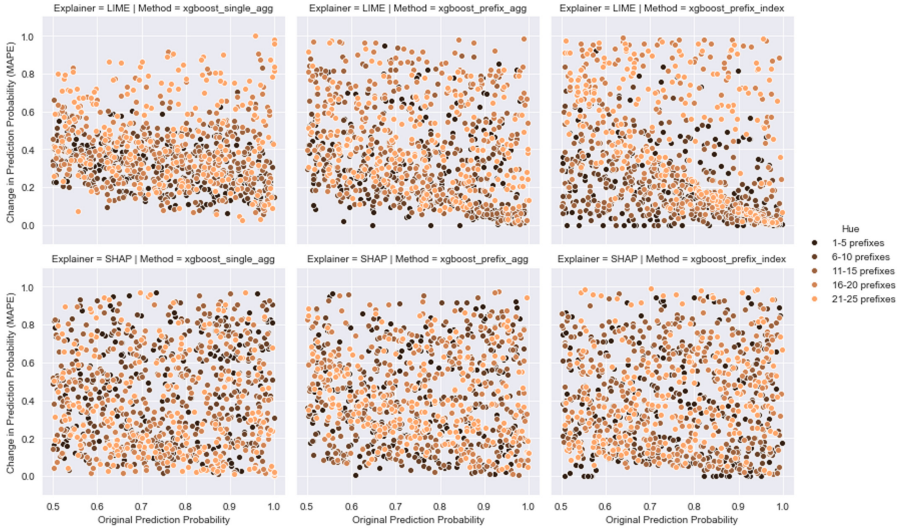


Fig. 1. Fidelity results over original prediction probability and prefix length of LIME and SHAP for the BPIC2012 dataset.

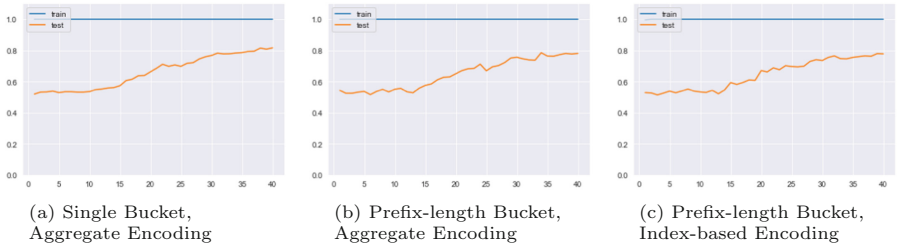


Fig. 2. Accuracy of predictive models at each prefix length for the BPIC2012 dataset

to the poor fidelity results. As noted, perturbation of the input was conducted using the feature vector, not the underlying event log, and potential dependencies between the features (such as events and their attributes) could also have contributed to these poor results. If this is true, it is possible that the combination of single bucketing with aggregate encoding produced the least faithful explanations as it preserves the complexity of event logs the least.

Secondly, the poor fidelity results may also be due to some internal mechanisms of the explanation methods. In particular, sampling methods used by LIME to produce surrogate models are known to often produce poor results [7]. It is possible that the underlying mechanisms of LIME and SHAP cannot appropriately recreate the complex dependencies between the features that can be derived from event logs.

Thirdly, it is likely that the characteristics of the underlying black box also affected the fidelity results. This is most noticeable in the BPIC2012 dataset,

when fidelity increased as the accuracy of the black-box model increased at higher prefix lengths. This phenomenon of increased accuracy closer to the completion of a running instance is expected in predictive process monitoring [10]. It is possible that the poor performance of the black-box models at smaller prefix lengths led to overall low fidelity and a poorly-fit surrogate model. Also, in predictive process analytics, an accurate prediction as early as possible during the process is valuable. Accordingly, any corresponding explanations will also perform better if they are faithful at an earlier stage during process prediction.

Insights for Future Work. Based on the above findings, we propose that a number of considerations should be made when assessing the fidelity of process prediction explanations. Firstly, a closer investigation needs to be made of the perturbation method used, with consideration of whether perturbation of the event log would be more appropriate than perturbation of the features extracted from it. Secondly, it would also be useful to consider the assumptions made when developing the proposed evaluation method, including the assumption that the boundaries derived from the explanation are absolute. Thirdly, to better understand whether the poor results stem from the use of event logs as datasets, the proposed method should be applied with more standard, less complex tabular datasets. Furthermore, evaluations should also be conducted with other types of black-box models, to understand how the underlying black box, including accuracy, contribute to the faithfulness of explanations.

5 Conclusion

As black-box models are often applied in predictive process analytics, explainability becomes necessary to help understand why certain predictions are made by the underlying predictive models. To determine the quality of explanations, we have proposed a method to assess explanations of process predictive models. The application of this proposed method to three event log datasets using a variety of bucketing and encoding techniques has provided observations regarding the faithfulness of process prediction explanations. Insights derived from the observations can be used to guide future refinement of the proposed method and evaluations as well as understanding of process prediction explainability.

References

1. Du, M., Liu, N., Yang, F., Ji, S., Hu, X.: On attribution of recurrent neural network predictions via additive decomposition. In: The World Wide Web Conference, WWW 2019, pp. 383–393. ACM (2019)
2. Galanti, R., Coma-Puig, B., de Leoni, M., Carmona, J., Navarin, N.: Explainable predictive process monitoring. In: 2nd International Conference on Process Mining, pp. 1–8. IEEE (2020)
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5), 93:1-93:42 (2019)

4. Kindermans, P.-J., et al.: The (un)reliability of saliency methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. LNCS (LNAI), vol. 11700, pp. 267–280. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_14
5. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: *Annual Conference on Neural Information Processing Systems*, pp. 4765–4774 (2017)
6. Messalas, A., Kanellopoulos, Y., Makris, C.: Model-agnostic interpretability with shapley values. In: *10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019*, pp. 1–7. IEEE (2019)
7. Rahnama, A.H.A., Boström, H.: A study of data and label shift in the lime framework (2019). [arXiv: 1910.14421](https://arxiv.org/abs/1910.14421)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: “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*, pp. 1135–1144 (2016)
9. Sindhgatta, R., Ouyang, C., Moreira, C.: Exploring interpretability for predictive process analytics. In: Kafeza, E., Benatallah, B., Martinelli, F., Hacid, H., Bouguettaya, A., Motahari, H. (eds.) *ICSOC 2020*. LNCS, vol. 12571, pp. 439–447. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65310-1_31
10. Teinemaa, I., Dumas, M., Rosa, M.L., Maggi, F.M.: Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Trans. Knowl. Discov. Data* **13**(2), 17:1–17:57 (2019)