Design and Evaluation of Explainable Methods for Predictive Process Analytics

Mythreyi Velmurugan[0000−0002−5017−5285]

Queensland University of Technology, Brisbane, Australia
mythreyi.velmurugan@hdr.qut.edu.au

Abstract. Predictive process analytics focuses on predicting the future states of running instances of a business process using machine learning and AI-based techniques. While advanced machine learning techniques have been used to increase accuracy of predictions, the resulting predictive models lack transparency. Current explainability methods, such as LIME and SHAP, can be used to interpret black box models. However, it is unclear how fit for purpose these methods are in explaining process predictive models, which use complex, multi-dimensional event logs, often alongside various types of context-related data. However, given the vast array of explainable methods available, and the differences in the mechanism used to provide the explanations and the content of the explanation, this evaluation becomes complex and no standard method or framework of evaluation currently exists. The proposed project, therefore, aims to address methods to evaluate and improve AI and machine learning transparency in the field of predictive process analytics, with particular emphasis on creating a standardised evaluation framework and approach for event log data.

Keywords: Predictive process analytics · Explainable AI · Transparency · Evaluation frameworks

1 Introduction

Business Process Management (BPM) methods are increasing in technical complexity and sophistication. An emerging BPM tool is predictive process analytics (PPA), which, at runtime, attempts to predict some future state of a process using machine learning models [11]. Modern data analytics techniques and increased availability of machine-generated process data has enabled PPA, which applies predictive analytics to business processes, providing a powerful tool that can be used to support better decision-making in organisations, optimisation or other process management activities [15].

However, the opaque nature of some prediction systems are cause for concern. While more complex and sophisticated prediction algorithms often produce more accurate predictive models, these models are also less transparent – an issue that could affect an organisation's transparency, ethical conduct, accountability and liability, and raise potential issues with the safety of and fairness towards stakeholders [5]. Methods and techniques have been proposed in machine learning to explain such opaque "black box" models, forming a research theme known as explainable AI (XAI) [5]. Several recent studies in PPA have applied existing

XAI techniques to interpret process prediction models (for example, in [4, 17, 18]). However, given the variety of explainable methods available and emerging in the field of XAI, it is unclear how fit for purpose any given intepretability technique is when applied to PPA. Frameworks and metrics for determining XAI fitness for PPA are so far under-explored, and further investigation is needed to understand the impacts of dataset and model characteristics on explanation quality. In particular, a standardised evaluation framework and approach are necessary, particularly for tabular data and sequential tabular data, such as event logs.

The proposed project, therefore, will address methods to improve AI and machine learning transparency for predictive process analytics, provide a framework for evaluating these explainable methods and attempt to provide a set of guidelines or recommendations for PPA explainability. There will be a focus on creating explanations to empower decision-making, with emphasis on explanation understandability and comprehensibility. This paper is structured as follows. Section 2 introduces PPA, explainable PPA and XAI in more detail. The research gaps to be explored by the proposed project, and the approach for the proposed research are explained in section 3, and progress achieved to date is presented in section 5. Planned future work is highlighted in section 6, and section 7 concludes this paper.

2 Background and Related Works

Process mining forms the backbone of PPA, where event data is extracted from historical event logs, and used by a prediction model to predict some future state of running process executions (known as the prediction target), such as how a running process will end, time to the end of the process, or future sequences of activities [11]. Most techniques for PPA require data processing, followed by learning and prediction, though there may be some variations depending on the data and learning algorithms used. To improve prediction accuracy, event log data can be supplemented by contextual information that might prove useful in making the final prediction, such as case documentation [11].

A two-phase approach is generally applied to PPA [20, 21], which begins with an offline processing and learning phase where the predictor is created and trained, followed by deployment of the predictor at runtime (see figure 1). The first phase begins with the processing and combining of historical event log data and contextual data to extract relevant process information. Generally, the relevant information, including activities that have been completed, attributes associated with each activity, and any contextual information, are grouped together by the case for which that activity occurred, creating a log of prefixes that describe a single run of the process – a single process instance. These prefixes then converted into features more appropriate to train a predictive model, and used to create a predictive model. This processing and conversion is also necessary in the second phase, when runtime data is used to predict some future state of a running process instance.

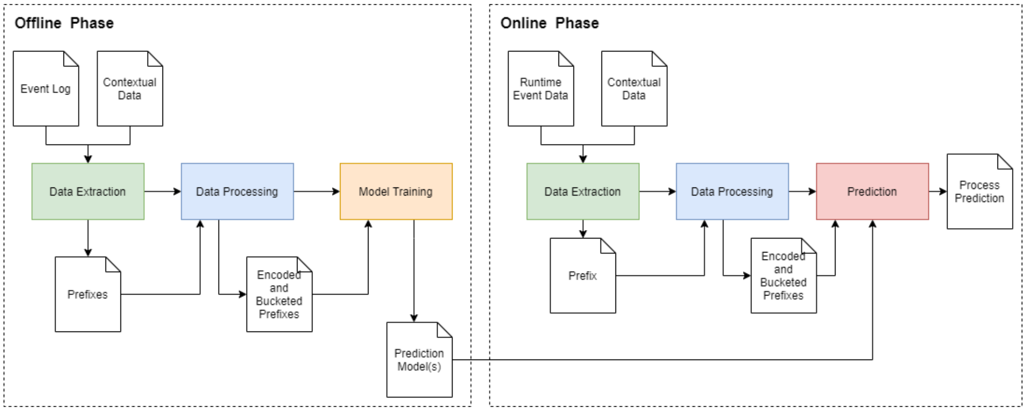

Fig. 1. Two-phase approach to process prediction

One complexity in predictive process analytics comes from the data used. Event log data alone is highly complex, containing various attributes for each activity in a case, including static attributes that do not change throughout a process instance, as well as dynamic attributes that often do change throughout the process instance, both of which must be encoded, though the degree to which the temporal information is preserved during encoding often varies. If combined with context data, which may or may not be temporal in nature, this encoded dataset becomes even more complex. Given this dataset complexity, and the inherently opaque nature of the machine learning algorithms that are most effective in creating accurate process predictions [18], it becomes hard to understand why a predictive model may have returned any given prediction. Furthermore, this lack of clarity leads to poorer understanding of model trustworthiness and impairs a user's ability to make informed decisions based solely on a predictive model's output [15].

As such, explainable predictive process analytics has emerged as an attempt to increase the transparency of these process predictive models. State-of-the-art explainable predictive process analytics have generally attempted to apply existing explainable methods to PPA. Interpretability in machine learning is generally broken down into two categories: interpretable prediction models and post-hoc interpretation. Interpretable prediction models are those that are generated in such a way as to be immediately interpretable by a human [5], though this often means that the models are simpler, and so may have reduced predictive power. The most common interpretation mechanism to be applied in explainable PPA is post-hoc interpretation, where an interpretation mechanism external to the predictive model is applied after the creation of the model. For example, the use of LIME and SHAP to evaluate and improve black box models [17, 18] has been explored, and SHAP has been used to create explainable dashboards for informed decision-making [4]. While this allows for the use of a more sophisticated predictive model, this does not necessarily imply than this external mechanism can accurately interpret the black box model. As such evaluation is necessary to understand the inherent fitness of such methods to interpreting black box predictive process models.

3 Research Objectives

While there have been attempts to apply or create XAI methods to interpret and explain process predictions, the fitness of these methods to explain PPA is unknown, and there are few frameworks available to evaluate this fitness. There are a number of ways to evaluate explanations, many of which depend on the purpose of the explanation and the explainee [13]. Although a decontextualised evaluation will not necessarily provide an indication of explanation usefulness or user satisfaction [8], it is still necessary to understand the inherent fitness of the explainable method for the prediction problem. As such, it is important to consider the characteristics necessary for an explanation and find suitable evaluation measures. Therefore, the proposed research problem for this project is: How can the fitness of explainable methods be assessed when explaining predictive process analytics? This can be broken down into the following research questions:

- **RQ1**: What AI-enabled models and techniques can be used to present explanations to users in the context of process prediction?

  Given the broad range and diversity of interpretation methods available, all intended for different purposes and functioning in different ways, it becomes necessary to identify interpretability methods and tools relevant to the datasets and methods used in PPA. Therefore, the state-of-the-art must be first understood, and XAI methods that are commonly applied to PPA or are applicable to PPA must be identified.

- **RQ2**: What criteria would be suitable to assess the quality of process prediction explanations generated from the models and techniques identified in RQ1?

  While a number of evaluation frameworks and taxonomies exist for evaluating interpretability and explainability tools, they are highly generalised and act as broad categorisations, rather than an evaluation method or standard. For example, in [19], while a number of dimensions of categorising and evaluating XAI are presented, there are no specific metrics and methods that can be applied for evaluation, nor any standardised frameworks or approaches. As such, it becomes necessary to define a suitable evaluation framework or approach for functionally-grounded evaluation of XAI, with emphasis on extensibility and flexibility to account for the different prediction problems, datasets, explanation types and users that may be involved.

- **RQ3**: What methods and approaches can be used to evaluate explainable methods for predictive process analytics, given the criteria RQ2 and the methods identified in RQ1?

- **RQ3.1:** What standard approaches and/or methods can be used to evaluate explanations created for prediction problems using tabular data?

- **RQ3.2**: How can standard approaches and/or methods used to evaluate explanations created for tabular data be adapted for event logs?

RQ3 is necessary in order to better understand how fit for purpose the identified XAI methods are for PPA. Firstly, functionally-grounded evaluation is necessary in order to determine how well-suited an XAI method inherently is to solving the problem of explaining PPA, even before users can be considered. However, many evaluation specific methods and metrics for XAI in literature are specific to a particular explainable method (such as in [22]), or unsuited for tabular data (such as those used in [2,3]). Therefore, a standardised and generalisable approach for tabular is necessary, before this can be adapted for sequential tabular data, such as event logs. The scope of the research required for this question will be determined by the relevant explainable methods identified in RQ1 and the specific criteria for determining quality determined by RQ2.

4 Research Methodology

Design Science Research (DSR) will be used to guide the methods used in the proposed project. Hevner et al. [7] define DSR as a problem solving paradigm that creates innovations that can be used to formalise practices, ideas and products, and in doing so facilitate the effective and efficient creation, usage and maintenance of information systems in businesses. In a later paper, Hevner [6] identifies three research cycles in DSR, which connect the three facets of environment, knowledge base and the research itself together in an iterative way.

- *The Relevance Cycle*, which grounds the research to a problem domain. Acceptance criteria for the output of the research is based on the problem domain, and the effectiveness of the output will be considered in the context of the problem domain. In the proposed project, the problem domain is predictive process analytics, and the ultimate outcomes of the project (approaches for determining the quality of explanations) will be evaluated within the context of process predictions.
- *The Rigor Cycle*, in which information is drawn from and added to a knowledge base during the course of research. In the proposed project, existing prediction and explainability methods and models will be used to create process prediction explanations, and existing approaches for evaluation will be studied. The new or adapted evaluation frameworks, approaches and methods that will be used to assess process prediction explanations, the results of the evaluations, as well as any new constructs created, will be added to the knowledge base.
- *The Design Cycle*: This cycle will form the major part of the proposed project, where evaluation approaches will be designed, explanations will be created and evaluation conducted.

The work will be conducted in three phases:

- **Phase One** will be comprised of the Problem Identification and Motivation and Objective Definition stages of the DSR methodology, wherein the interpretability and explanation needs of PPA and relevant explanation methods will be explored (RQ1) and the characteristics required for PPA explanations and ways to determine the fitness of explainable methods for PPA will be defined (RQ2).
- **Phase Two** will be an iteration the Design and Development, Demonstration and Evaluation stages of the DSR methodology that attempts to answer RQ3.1. In this phase, standard approaches and methods for evaluating explanations of tabular data will be developed, assessed and refined.
- **Phase Three** will be another iteration of the Design and Development, Demonstration and Evaluation stages, this time with the aim of answering RQ 3.2. In this phase, the standard approaches and methods created during Phase Two will be adapted for event logs, assessed and further refined as necessary.

Results of the research will be communicated throughout all three phases. It is expected that the following outcomes will be achieved as a result of this project:

1. Proposal of evaluation criteria for evaluating explanations and explanation methods for process predictions;
2. Proposal of a standard approach to evaluating explanations for tabular data;
3. Proposal of specific evaluation methods for classes of explainable methods used to explain process predictive models that are generalisable within those classes; and
4. Functionally-grounded evaluation of several existing explainable methods to determine their fitness for explaining PPA;

5 Progress To Date

A three-level system of evaluation that considers context to differing levels is proposed in [1], which comprises of:

- *Application-Grounded Evaluation*: Evaluating explanations in full context with end users;
- *Human-Grounded Evaluation*: Evaluating explanations with laypeople doing simple or simplified tasks; and
- *Functionally-Grounded Evaluation*: Using functional tasks to evaluate without subjective, user-based evaluation.

This PhD will primarily focus on the first level of functionally-grounded evaluation. Functionally-grounded evaluation measures include those that do not require human input, but still allow some objective judgement of the explanation to be made. A common, and important, example of such a measure would be *computational efficiency* [13]. This is particularly important in PPA, where the earliness of interventions in a process instance affects ability to change undesirable outcomes. Another key requirement for explanations is consistency, also known as *explanation stability*. This is defined as the level of similarity between explanations for similar or identical instances [22], and is vital in ensuring trust in the explanation and the predictive model, and ensuring actionability. Also vital is the accuracy of the explanation with respect to the original predictive model, known as *explanation fidelity*. This is generally defined as how faithful the explanation is to the black box – that is, how accurately the external mechanism of the explainable method mimics the black box.

An initial functionally-grounded evaluation was conducted with three event log datasets and two explainable methods for the outcome prediction problem, where the final outcome of any given case was being predicted. The two explainable methods chosen for initial testing were Local Interpretable Model–Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). These two methods are currently the most prevalent explainable methods in PPA literature (for example, used in [4, 17, 18]), and are also among the most well-known and popular methods in XAI. As such, they were considered suitable for initial evaluations. LIME and SHAP both offer local feature attribution explanations [9, 16]. That is, for each instance (input) requiring explanation, these explainable methods will rank the features in the input in order of importance and provide a weight to each feature to signify its overall contribution to the black box model's final prediction.

A key challenge for this initial evaluation was, and still is, the lack of standardised or generalisable methods that can be applied for evaluation and comparison of the two explainable methods, particularly explanation stability and fidelity. Given their relative popularity in the field of XAI, it was assumed that existing evaluation methods would exist for LIME and SHAP that can be adapted for event log data, and later for other explainable methods. Although event log data are complex, and include a time-series component that sets a particular sequence to the features present within the data, event logs are similar in construction to tabular data. Moreover, when processed for machine learning algorithms, this data is often formatted in such a way that temporal information is only partially preserved, and the final input to the model is identical in format to that of standard tabular data. As such, the initial plans for testing relied on using existing evaluation methods for LIME and SHAP, and adapting these methods to consider the temporal aspect if necessary. However, existing evaluation methods and metrics are often specific to a particular explainable method [22] or data types other than tabular, such as images or text [2, 3]. As such, an evaluation method that would allow comparison of the two methods had to be developed for the initial testing based on existing methods.

In order to develop such methods, existing approaches were adapted. In particular, evaluation methods and evaluation metrics used to assess the stability of feature selection algorithms are being applied to test the stability of LIME and SHAP. Much as LIME and SHAP do, feature selection algorithms provide a ranking or subset of the most useful features to use in evaluation, and also often offer some kind of "feature weight" to indicate the importance of each feature, and the stability of all three may be assessed [12,14]. As such, it was decided that explanation metrics used to assess the stability of feature selection algorithms can be adapted to evaluate the stability of LIME and SHAP. In particular, two types of stability are being assessed for feature attribution explainable methods used: *stability by subset*, which assesses the stability of the subset of the most important features; and *stability by weight*, which assesses the stability of the weight assigned to each feature as part of the explanation.

The evaluation method for explanation fidelity has, thus far, proven to be more complex to develop. The initial approach mimicked a commonly-used ablation approach for assessing the fidelity of image and text data, wherein the features determined to be relevant by the explanation are removed and the change in the prediction probability for the original prediction is used to determine the correctness of the explanation [2, 3]. While this removal is relatively simple in other types of datasets, models built on tabular data automatically impute values into "gaps" in the data, or assign missing values to be the equivalent of an infinite value. As such, this ablation approach was replaced with a perturbation approach, where the values of the features deemed relevant by explanations were altered to inject noise into the input data. The results of this evaluation, for both LIME and SHAP, were quite poor and further investigations are currently being conducted to further evaluate and adjust this perturbation approach as necessary.

6 Future Work and Challenges

There are a number of further activities that need to be undertaken to fully answer the research questions outlined. Currently, evaluations have been conducted with LIME and SHAP, which are feature attribution methods – local methods that explain the effects of different components for a single prediction [10]. These two methods alone are not representative of the vast array of explainable methods currently available, and as such approaches that can be used to assess other classes of explainable methods must also be considered. Of particular interest are explainable methods specific to time series data.

A key challenge here will be to design explainable method evaluation methods in order to ensure comparability between explainable methods of different classes. For example, one type of stability currently measure is the stability of weights assigned to each feature. If techniques other than feature attribution are to be evaluated, this measure will no longer be relevant, but other types of stability may become more relevant, such as the stability of predicates in rule-based explanation. However, the results produced by the evaluations must still be comparable while being suitable for the relevant explanation methods, in order to enable the ability to compare and choose between explainable methods given a predictive problem. The ultimate goals of this project will be to create a functionally-grounded evaluation approach to assess the suitability of any given explainable method in the relevant classes of explainable methods for time-series based prediction problems, such as PPA.

Furthermore, the evaluation methods and approaches that will result from this PhD will also also require validation and evaluation. A viable way of doing so may be to use simpler, inherently interpretable and transparent machine learning models when developing the evaluation methods, and confirm the validity of the methods against a transparent model. Existing methods and measures from other, similar fields, are also being adapted as necessary, as is the case with the adaptation of stability measures and methods from the field of feature selection.

7 Conclusion

Post-hoc explainable methods are gaining popularity as a means of improving the transparency of process predictive models. However, the fitness of these methods for predictive process analytics is extremely unclear. Given that no standard approaches for evaluation of explainable methods exist, particularly for process prediction explanations, the project outlined in this document will attempt to propose evaluation criteria for evaluation explanations and explainable methods for process predictions; create standardised approaches for evaluating explainable methods for process predictions; and propose evaluation methods for classes of relevant explainable methods with regards to explaining process predictions.

Acknowledgements I would like to thank my supervisors Dr Chun Ouyang, Dr Catarina Moreira and Dr Renuka Sindhgatta for their ongoing support and guidance.

References

1. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning, arXiv: 1702.08608v2
2. 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 '19. ACM Press (2019). https://doi.org/10.1145/3308558.3313545
3. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, Italy (22-29 October 2017). https://doi.org/10.1109/iccv.2017.371
4. Galanti, R., Coma-Puig, B., de Leoni, M., Carmona, J., Navarin, N.: Explainable predictive process monitoring, arXiv: 2008.01807
5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys **51, Article 93**(93) (Jan 2018). https://doi.org/10.1145/3236009
6. Hevner, A.R.: A three cycle view of design science research. Scandinavian Journal of Information Systems **19**, 87–92 (2007), https://aisel.aisnet.org/sjis/vol19/iss2/4
7. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**, 75–105 (2004). https://doi.org/10.2307/25148625
8. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for Explainable AI: Challenges and Prospects (2018), arXiV:1812.04608v2
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 2017 Neural Jnformation Processing Systems Conference. Long Beach, USA (4-9 December 2017)

10. Maksymiuk, S., Gosiewska, A., Biecek, P.: Landscape of r packages for explainable artificial intelligence (Sep 2020), arXiv: 2009.13248
11. Marquez-Chamorro, A.E., Resinas, M., Ruiz-Cortes, A.: Predictive monitoring of business processes: A survey. IEEE Transactions on Services Computing **11**(6), 962–977 (Nov 2017). https://doi.org/10.1109/tsc.2017.2772256
12. Mohana Chelvan, P., Perumal, K.: A Survey of Feature Selection Stability Measures. International Journal of Computer and Information Technology **5**(14), Article 14 (2016), https://www.ijcit.com/archives/volume5/issue1/Paper050114.pdf
13. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems, arXiv: 1811.11839v4
14. Nogueira, S., Sechidis, K., Brown, G.: On the Stability of Feature Selection Algorithms. Journal of Machine Learning Research **18**(174), Article 174 (2018), http://jmlr.csail.mit.edu/papers/volume18/17-514/17-514.pdf
15. Rehse, J.R., Mehdiyev, N., Fettke, P.: Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory. Künstliche Intelligenz **33**(2), 181–187 (Apr 2019). https://doi.org/10.1007/s13218-019-00586-1
16. 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. San Franciso, California (13-17 August 2016). https://doi.org/10.1145/2939672.2939778
17. Rizzi, W., Francescomarino, C.D., Maggi, F.M.: Explainability in predictive process monitoring: When understanding helps improving. In: International Congerence on Business Process Management 2020. pp. 141–158. Springer International Publishing, Seville, Spain (13-18 September 2020). https://doi.org/10.1007/978-3-030-58638-6_9
18. Sindhgatta, R., Ouyang, C., Moreira, C.: Exploring interpretability for predictive process analytics. In: 18th International Conference on Service-Oriented Computing. LNCS, vol. 12571, pp. 439–447. Springer (2020)
19. Sokol, K., Flach, P.: Explainability fact sheets. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain (27-30 January 2020). https://doi.org/10.1145/3351095.3372870
20. Teinemaa, I., Dumas, M., La Rosa, M., Maggi, F.M.: Outcome-oriented predictive process monitoring: review and benchmark. ACM Transactions on Knowledge Discovery in Data **13, Article 17**(17) (Jun 2019). https://doi.org/10.1145/3301300
21. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Teinemaa, I.: Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM Transactions on Intelligent Systems and Technology **10, Article 34**(34) (Aug 2019). https://doi.org/10.1145/3331449
22. Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D.: Statistical stability indices for lime: obtaining reliable explanations for machine learning models, arXiv: 2001.11757v1