Challenges Towards Production-Ready Explainable Machine Learning

Lisa Veiber, Kevin Allix, Yusuf Arslan, Tegawendé F. Bissyandé, and Jacques Klein, *SnT – Univ. of Luxembourg*

https://www.usenix.org/conference/opml20/presentation/veiber

This paper is included in the Proceedings of the 2020 USENIX Conference on Operational Machine Learning.

July 28–August 7, 2020

978-1-939133-15-1

Open access to the Proceedings of the 2020 USENIX Conference on Operational Machine Learning is possible thanks to the generous support of

NetApp®

Challenges Towards Production-Ready Explainable Machine Learning

Lisa Veiber
SnT – Univ. of Luxembourg

Kevin Allix
SnT – Univ. of Luxembourg

Yusuf Arslan
SnT – Univ. of Luxembourg

Tegawendé F. Bissyandé
SnT – Univ. of Luxembourg

Jacques Klein
SnT – Univ. of Luxembourg

Abstract

Machine Learning (ML) is increasingly prominent in organizations. While those algorithms can provide near perfect accuracy, their decision-making process remains opaque. In a context of accelerating regulation in Artificial Intelligence (AI) and deepening user awareness, explainability has become a priority notably in critical healthcare and financial environments. The various frameworks developed often overlook their integration into operational applications as discovered with our industrial partner. In this paper, explainability in ML and its relevance to our industrial partner is presented. We then dis- cuss the main challenges to the integration of explainability frameworks in production we have faced. Finally, we provide recommendations given those challenges.

1 Introduction

The increasing availability of data has made automated techniques for extracting information especially relevant to businesses. Indeed, AI overall contribution to the economy has been approximated to $15.7tr by 2030 [8]. State-of-the-art frameworks have now outmatched human accuracy in complex pattern recognition tasks [6]. However, many accurate ML models are—in practice—black boxes as their reasoning is not interpretable by users [1]. This trade-off between accuracy and explainability is certainly a great challenge [6] when critical operations must be based on a justifiable reasoning.

Explainability is henceforth a clear requirement as regulations scrutinises AI. In Europe, our industrial partner faces GDPR, the General Data Protection Regulation, which includes the *right to explanation*, whereby an individual can request explanations on the workings of an algorithmic decision produced based on their data [5]. Additionally, user distrust, from their lack of algorithmic understanding, can cause a reluctance in applying complex ML techniques [1, 2]. Explainability can come at the cost of accuracy [10]. When faced with a trade-off between explainability and accuracy, industrial operators may, because of regulation reasons, have to resort to less accurate models for production systems. Finally, without explanations, business experts produce by themselves justifications for the model behaviour. This can lead to a plurality of explanations as they devise contradicting insights [3].

Integrating explainability in production is a crucial but difficult task. We have faced some challenges with our industrial partner. Theoretical frameworks are rarely tested on operational data, overlooking those challenges during the design process. Overcoming them becomes even more complex afterwards.

2 Explainability

Explainability is rather ill-defined in the literature. It is often given discordant meaning [7] and its definition is dependent on the task to be explained [4]. Nonetheless, we use explainability interchangeably with interpretability [7] and define it as aiming to respond to the opacity of the inner workings of the model while maintaining the learning performance [6].

From this definition, explainability can be modelled in different ways. First, interpretability can be either *global* or *local*. Whereas the former explains the inner workings of the model at the model level [4], the latter reaches interpretability at the instance level. Furthermore, there is the distinction between inherent and post-hoc interpretability. Inherent explainability refers to the model being explainable [1], while post-hoc explainability entails that once trained, the model has to further undergo a process which will make its reasoning explainable [9, 10]. This results in four different types of explainability frameworks. The effectiveness of the frameworks depends on the type chosen with respect to the task of the model we want to explain.

Moreover, explainability is usually achieved through visualization-based frameworks, which produce graphical representations of predictions, or text explanation of the decision [1]. Effective visualizations or textual description with a decision can be sufficient to reach explainability [3]. Still, those supposedly-interpretable outputs are rarely validated through user studies. In our case, the frameworks, which were not validated in such a way, yielded models that have proven to be just as non-interpretable as the original ML model for our industrial partner.

Various reasons for explainability were previously mentioned. Our industrial partner was further interested in explainability for audit purposes as they must provide justifications for the automated system decisions. Besides, they were concerned with potential biases within training datasets. Some features, such as gender, although perhaps appearing as effective discriminators in ML models, cannot be legally used in business operations analytics [12]. Yet, other less explicit attributes may be direct proxies to such restricted features, eventually creating biases in the models. Adding explainability to existing models can uncover existing biases arising from proxies included in the model. This allows the operator to change models when bias is identified.

3 Challenges to Explainability

3.1 Data Quality

Our industrial partner was implementing ML model on tabular data. One of the first challenge identified was that most frameworks are designed for Natural Language Processing and Computer Vision tasks. Thus, there are fewer frameworks focusing on tabular data. With this omission, complications relating to tabular data quality are not properly addressed. This resulted in limitations of the framework explainability. For instance, visualizations designed for continuous features became inefficient for categorical variables. Furthermore, frameworks tested on tabular data often rely on datasets which are optimal for visualization purposes. Our data was not optimal as it contained missing values, and clusters between classes were not clearly separated. For example, while the several approaches we tested reported impressive results on a selection of domain, it is often impossible to know beforehand whether or not it can be applied to a specific domain. Indeed, they proved to be far less valuable when applied to the real datasets of our partner. For one specific problem, we obtained visualizations such as shown in Figure 1. In this case, the display of the gradient projection of data points [10], which relies on comparing the different classes according to pairs of variables, was cluttered. There was no clear distinction between the classes, hence the framework did not provide interpretability.

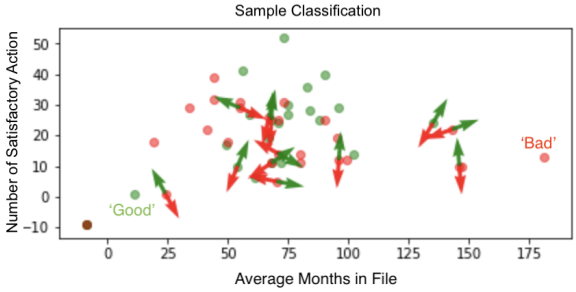

Figure 1: Example of visualization-based interpretability framework [10]

3.2 Task and Model Dependencies

Our industrial partner was implementing Random Forests. However, some frameworks are designed for specific model types, more particularly for Neural Networks and Support Vector Machines. This has been addressed by model agnostic approaches [9, 10] and surrogate models. Still, the latter loses accuracy when approximating the unexplainable model, only providing explanations when both models reach the same predictions, but not when they differ. Moreover, explainability is also task-dependent. Our industrial partner needed different explanations for different audiences. Yet, we detected through our experiments an emphasis on theoretical review of task dependency rather than empirical research. This insufficiency of practical consideration limits frameworks deployment in production, as a lengthy experimental process is required to inspect which explanations best fit the task.

3.3 Security

Another challenge from explainability in production is security. This was a significant concern for our partner. Indeed, if clients can have access to the reasoning of the model decision-making, they could apply adversarial behaviours by incrementally changing their behaviour to influence the decision-making process of the model. Thus, explainability raises robustness concerns preventing its immediate deployment in production. Furthermore, in a recent paper [11], it was shown that under strong assumptions, an individual having access to model explanations could recover part of the initial dataset used for training the ML algorithm. Given the strict data privacy regulations, this remains a case to investigate.

Implementing explainability frameworks in production can therefore significantly slow and complicate the project.

4 Conclusion

Data is crucial to derive information on which to base operational decisions. However, complex models achieving high accuracy are often opaque to the user. Explainable ML aims to make those models interpretable to users. The lack of research on operational data makes it challenging to integrate explainability frameworks in production stages. Several challenges to explainability in production we have faced include data quality, task-dependent modelling, and security. Given those challenges we recommend industrials to (1) clearly define their needs to avoid obstacles defined are task and model dependencies in this paper, (2) give more consideration to possible industrial applications when frameworks are designed, (3) undertake systematic user validation of frameworks to evaluate the explainability potential of those frameworks, (4) regarding the security challenges, we suggest to simulate adversarial behavior and observe the model behaviour to raise any robustness issues, (5) industrials could also undertake the exercise of recovering the entire training datasets given explanations and a small part of the original data.

References

[1] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, 2017.

[2] Chris Brinton. A framework for explanation of machine learning decisions. In *IJCAI-17 workshop on explainable AI (XAI)*, pages 14–18, 2017.

[3] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable AI really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

[4] Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2, 2017.

[5] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[6] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

[7] Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, June 2018.

[8] PwC. Pwc's global artificial intelligence study: Sizing the prize. https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html, 2017. Accessed Feb 2020.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. 2017.

[11] Reza Shokri, Martin Strobel, and Yair Zick. Privacy risks of explaining machine learning models. *arXiv preprint arXiv:1907.00164*, 2019.

[12] Naeem Siddiqi. Intelligent credit scoring: Building and implementing better credit risk scorecards. 2017.