Crimson Publishers
Wings to the Research

Short Communication

The Value of Explanations for Machine Learning Algorithms

Stockem Novo A*

Institute of Computer Science, Ruhr West University of Applied Sciences, Germany

ISSN: 2832-4463

***Corresponding author:** Stockem Novo A, Institute of Computer Science, Ruhr West University of Applied Sciences, 45479 Mülheim an der Ruhr, Germany

Submission: May 24, 2022

Published: July 06, 2022

Volume 2 - Issue 2

How to cite this article: Stockem Novo A. The Value of Explanations for Machine Learning Algorithms. COJ Rob Artificial Intel. 2(2). COJRA. 000533. 2022.
DOI: 10.31031/COJRA.2022.02.000533

Abstract

Machine Learning techniques are powerful in many different domains. For application in sensitive areas where humans are involved, the requirements regarding model understanding are strict. Currently, methods are developed that help understanding and allow drawing conclusions from the observations. Decision plots and counterfactual explanations give information about the model output, the impact of single features can be estimated. This is already a beneficial first step towards model transparency. However, there are no methods yet that use this gained knowledge as a feedback to update the model accordingly in a straight-forward manner.

Keywords: Deep learning; Bias; Fairness; Explainable AI

Introduction

Deep Learning models outperform classic approaches for many applications. Such models, often generalized under the term Artificial Intelligence (AI), shift the need for expert knowledge to a different domain: As for classical models, based on rules or simple often quasi-linear models, a proper understanding of the specific Use Case and excellent programming skills are required, Deep Learning models can be easily applied even by amateurs with the help of Auto ML methods. A large-scale dataset is fed into the system and results are obtained straight-forward without going too deep into the analysis of the data. This is a fact that needs to be handled with caution when bringing a model into deployment. There is a trend towards more and more complex models. This is made possible due to the permanently increasing computer capacity and a strong community pushing the development of open source frameworks. Models do not need to be limited to specific Use Cases any more. For example language models can handle different languages or give similar performance for different technical domains. The recent trend of multi-task learning [1] is showing great success and providing astonishing results.

Taking into account all the benefits that such generalized models can bring, this comes at the cost of little understanding of how a model comes to its conclusions. This aspect is an active field of research under the term Trusted AI. The FAT model [2] requests an understanding of the essential parts fairness, accountability and transparency. In a top-down approach, it tries to attain a basic understanding of any Machine Learning Model (ML model). Fairness relates to the ML model output. Different groups must have equal opportunities or chances. Accountability requests for an entity which takes responsibility of the model decisions. It furthermore concerns the aspects reproducibility, robustness as well as safety and security. Model transparency addresses the technical level by asking for explanations of the model output.

Whenever a model affects people's life's, fairness needs to be considered critically. One essential problem source of fairness is bias. Bias by itself basically means an imbalance and does not need to consequently harm. However, if bias is prevalent in the dataset, the model will learn a representation of this imbalance and this might affect different groups.

As a consequence one should distinguish between dataset bias and model bias. Model bias can occur even if there is no bias in the dataset. Since this is state-of-the-art research, there are currently several methods being developed for explaining AI decisions and revealing sources of bias. We will discuss some prominent methods in the next section and draw a conclusion based on a Case Study for People Analytics.

Explaining AI Decisions

When asking for explaining a model, the central question is how the input relates to the final model decision and why. The mathematics is deterministic even for black-box models like Artificial Neural Networks and thus comprehensible. What is not clear is the choice of the parameters, i.e., the weights of the network, which are a reflexion of the feature importance. There are a number of explanation methods which can help understanding the importance of individual features: The method SHAP (Shapley Additive exPlanations) [3] is based on the idea of Shapley values coming from cooperative game theory. It estimates the importance of a single feature on the model result. For that, different feature subsets are drawn. The model is then computed with and withholding that feature and the model metric is weighted over different feature combinations resulting in the SHAP value of a feature. LIME (Local Interpretable Model-Agnostic Explanations) [4] generates new data points in the vicinity of a data point and uses them to approximate a linear surrogate model, which can be easily understood. Further explanation methods use decision boundaries to deduce rules. All these methods help understanding the influence of a feature on the model output and are thus helpful for providing transparency. When bringing a model into deployment, it is recommended to provide model cards [5] containing details about the model performance and known limitations. As a new trend, recently also data cards have been introduced [6].

Case Study: People Analytics

Consider a data-driven model for identifying employees who should receive a training in order to increase performance. Such a classification model can be trained on historical company data. Possible features are:

A. Gender

B. Age

C. Years of working for the company

D. Number of projects

E. Number of work packages

F. Time per project work package

The target variable shall be the binary variable: Received training (yes/no).

It is clear that this Use Case is subject to unfair treatment of different groups. This may affect gender groups male/female/diverse, different age groups, people of specific demographic background or others. In order to avoid a model bias, any bias in the dataset regarding these different groups shall be removed. This can be done by applying a preprocessing method, e.g. reweighing of the model input, during training as in-processing method or as post-processing method by e.g. transformation of the model outputs [7]. Still, it is possible that a model bias is learned that leads to unfair treatment of a group. For example, new employees could have a reduced work load, such that there is an intrinsic representation of the variable „Age“ in „Number of work packages“. Such dependencies can be more subtile and not easy to identify. In this case, the above mentioned explanation methods can be used to explain why the model came to a decision by evaluating the contribution of different features. Currently, the consequence of such an observation, is only a retraining of the model with different conditions, i.e., different feature sets or a change in the hyperparameters. There is little possibility of taking direct influence on changing a specific outcome based on expert knowledge.

Conclusion

Current explanation methods help understanding why a decision was made by a model. This helps bringing transparency to data-driven approaches. However, when a specific outcome is not acceptable, the problem can be solved only in a re-training or a fine-tuning of a model. Currently, the explanation methods have no direct feedback method that changes the model in accordance with expert knowledge.

References

1. Zhang Y, Yang Q (2021) A survey on multi-task learning. In: IEEE Transactions on Knowledge and Data Engineering, pp. 1.
2. Shin D, Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior 98: pp. 277-284.
3. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems pp. 4765-4774.
4. Ribeiro MT, Singh S, Guestrin C (2016) 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.
5. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, et al. (2019) Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency pp. 220-229.
6. Pushkarna M, Zaldivar A, Kjartansson O (2022) Data cards: Purposeful and transparent dataset documentation for responsible AI. arXiv preprint arXiv:2204.01075.
7. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems pp. 3315-3323.