Interpretability in Machine Learning: concepts and challenges

Welemhret Welay, Baraki,
School of Informatics, University of Skövde
Högskolevägen 1, 541 28 Skövde, Sweden
a20welba@student.his.se , weldie.ed@gmail.com

Abstract*—Machine learning models have been able to present promising results in many fields like medicine, industries, and space science. However, most of the machine learning models lack interpretability of how they are working, how data manipulation and prediction are conducted. With the increasing focus on interpretable machine learning models, many interpreting methods are proposed.*

Articles focused on the Interpretability of Machine learning models were reviewed. This defined Interpretability in different methods and frameworks. The Predictive, Descriptive, and Relevancy (PDR) framework in the data science life cycle was discussed in detail with the two common interpretability categories: model-based and post hoc interpretability. Interpretation desiderata and *challenges were addressed. Finally, the* researcher's *reflections and thoughts, and conclusion were forwarded.*

Index Terms— Interpretability, desiderata, Machine Learning

I. Introduction

MACHINE LEARNING (ML) models are receiving attention in different fields. Their prediction capabilities are increasing. The complexity (Lipton, 2016) of ML models is increasing from easily interpretable (i.e. decision tree - decision rules) models to complex black-box models (i.e. deep neural networks). Interpreting the ML model's prediction results, algorithmic structures, and data manipulations is a growing area of research. This report focuses on the literal meaning of ML interpretability. From this report, the reader will get a better understanding of the concepts related to interpretable ML.

According to the Cambridge English Dictionary the word "interpretable"[1] means "possible to find meaning in a particular context". According to the Murdoch et al. (2019) to interpret means to extract information from data to give insights and knowledge. Doshi-Velez & Kim (2017) state that interpretability is the ability to explain in understandable terms to end-users. Lipton (2016) presents a framework that outlines both outputs of interpretable machine learning models and how interpretable models be achieved. Doshi-Velez & Kim (2017) stated how the interpretability methods can be evaluated.

Machine Learning models use objective functions most often as accuracy-based metrics (predictive accuracy)(Murdoch et al., 2019).

Lipton (2016) describes interpretability as a universally known fact, however ill-defined, and interpreting models exhibit quasiscientific character. Interpretability (Lipton, 2016) can be seen in the lens of the legal notion of the right to explanations. Athey and Imbens (2015) suggest that interpretability is the method of revealing an association between an interpretable model and data.

As stated by Lipton (2016) interpretability has no formal technical definition and meaning. In Murdoch et al. (2019), they frame interpretability as the use of ML models to extract information and knowledge from a collection of data. The relevant knowledge (Lipton, 2016; Murdoch et al., 2019) extracted from data using ML models can be produced using visualizations, natural language, mathematical models, and explain by example.

The concept of interpretability in Machine Learning is not well defined to enable researchers and practitioners to understand distinguishable areas within the field. The concept of interpretability (Murdoch et al., 2019) in Machine Learning refers to more than one concept. The confusion in defining the concept of interpretability will introduce divergence of real-world and Machine Learning problem formulations.

Interpretability can be the source of discrimination in the future livelihood of modern society. Because, our activity, day-to-day life is more dependent and related to technology, algorithms, and machine learning technologies that we are forced to use in our professional and personal lives.

In the next sections of this article, the Interpretability framework, desiderata of interpretability, evaluation methods, reflections, and thoughts were discussed in detail.

II. Interpretations in the Data Science Life Cycle

As stated by NICK HOTZ (2021), Data Science Life Cycle is an iterative set of a data science project development phases to deliver quality products. The data science life cycle structures with the problem domain, data collection, modeling, and deployment. Interpretability of ML models should have to be integrated and framed within the wider process of the data science project development life cycle.

There are two main interpretability categories (Lipton, 2016; Murdoch et al., 2019): Model-based interpretability is used when the underlying relationship is relatively simple like in decision trees, and post-hoc interpretability is used in complex black-box models like deep neural networks.

Interpretations can be produced as visualizations, natural language, mathematical equations, and explaining by example (Lipton, 2016).

Interpretations in the data science life cycle can be viewed using the conceptual framework of PDR (Murdoch et al., 2019) which stands for Predictive Accuracy, Descriptive Accuracy, and Relevancy.

[1] Interpretable https://dictionary.cambridge.org/dictionary/english/interpretable (accessed on Oct 7, 2021)

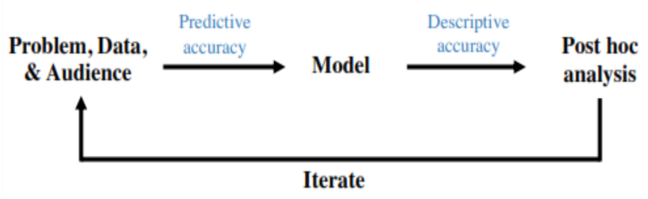

Figure 1 Predictive Accuracy, descriptive Accuracy, and Relevancy for Interpretability of ML models [Source: from (Murdoch et al., 2019)*]*

A. *Predictive accuracy, Descriptive accuracy, and Relevancy*

The Predictive, Descriptive, and Relevancy (PDR) as in the above Figure 1 is used to frame and guide the interpretability of ML models with the measures of predictive and descriptive accuracy. The predictive accuracy (Murdoch et al., 2019) is a measure of the ability to fit the data to the Machine learning model. The predictive accuracy (Murdoch et al., 2019) measures are quantifiable and easy to interpret. The descriptive accuracy has been defined by Murdoch et al. (2019) *as* "*the degree to which an interpretation method objectively captures the relationships learned by machine learning models*". Descriptive accuracy is used to measure the post hoc interpretation of machine learning models.

The relevancy (Murdoch et al., 2019) is measured by the users/audiences which is a subjective measure of how much the interpretable models are important in providing knowledge to the end-users. A conflict can happen in achieving both high predictive and descriptive accuracy E.g. in image analysis the model can achieve high predictive accuracy and can result in lower descriptive accuracy. In this sense relevancy (Murdoch et al., 2019) is important if it provides insight that solves the tradeoff in predictive and descriptive accuracy. Most often relevancy is the key to solving the tradeoff of predictive and descriptive accuracy.

The process in Figure 1 iterates until sufficient predictive accuracy (model-based interpretability) and descriptive accuracy (post-hoc analysis) are obtained. Finally, if there is a conflict in the two accuracies relevancy will resolve in a reasonable scenario. Within the framework of PDR model-based interpretability and post-hoc analysis will be discussed in the next section.

B. *Model-based interpretability*

Researchers develop machine learning models by collecting the required data and defining the required model parameters.

Model-based interpretability (Murdoch et al., 2019) and Transparency (Lipton, 2016) are used to refer to the same concept to understand the mechanism of how does the model works. The model-based interpretability (Lipton, 2016) has three main properties: simulatability, decomposability, and algorithmic transparency. Simulatability (Lipton, 2016; Murdoch et al., 2019) is defined as to reason about how its entire decision-making process takes place intuitively.

Modularity (Murdoch et al., 2019) or decomposability (Lipton, 2016) when some part of the decision process can be easy to compose and interpret. Algorithmic transparency (Datta et al., 2016) can also be concentrated on the algorithmic implementation of the underlying model in which how user data is manipulated and used.

C. *Post hoc interpretability*

At this stage, the practitioner (Lipton, 2016; Murdoch et al., 2019) can analyze a trained model in order to provide insights on the learned relationships at the post hoc analysis stage of the data science life cycle. In the posthoc analysis (Murdoch et al., 2019) there are two main interpretations: dataset-level (global) and prediction-level (local) interpretation by determining interaction and feature importance, and visualization. Additionally, textual explanations, identifying influential data points, and analyzing nearest neighbors are alternative methods in the post hoc interpretation.

III. Desiderata of Interpretability

As Lipton (2016) explained, end-users demand interpretability, that can't be captured by predictive accuracy (ML Objective function). Algorithms used in court and other personal data processing systems should have to simultaneously consider ethical, legality, and productivity considerations. These considerations (desiderata) are not handled with prediction accuracy. The desiderata discussed in (Lipton, 2016) are explained in detail below:

A. *Trust*

Interpretability (Lipton, 2016; Murdoch et al., 2019) can be described as a requirement for human beings to trust the machine learning model. Trust (Lipton, 2016) indicates confidence that machine learning models perform well with respect to real-world scenarios.

B. *Causality*

The cause and effect association (Lipton, 2016) of input and target variables is not reflected in the predictive accuracy of machine learning models. An Interpretable model should have to be visualized to show the associations of variables so that scientists can generate hypotheses to test the model in real-world problems.

C. *Transferability*

Human beings are significantly better at generalizing than ML models. Interpretability allows us to see how models perform when tested in a different setting than they were trained in. The Interpretability (Lipton, 2016) of the machine learning model should have to consider and test the machine learning model on how it behaves if the environment changes (Transferability). Interpretable models can be tested in different hypothesis scenarios in the real world.

D. *Informativeness*

As stated by Lipton (2016), the Interpretable machine learning model, which supports decision-makers expected to be informative on suggesting the possible decisions to the human experts. E.g. a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision.

E. *Fair and Ethical Decision-Making*

To determine if decisions did by models can conform to ethical standards, the user should have to interpret the model results. A new European Union proposed that individuals affected by algorithms have the right to an explanation. Algorithm decisions should have to be explainable and modifiable in case it is incorrect.

IV. Evaluation methods of Interpretable models

Measuring predictive accuracy (Murdoch et al., 2019) is straightforward and is simple to use. The most challenging part of evaluating interpretation in the PDR (Murdoch et al., 2019) framework is quantifying descriptive accuracy and relevancy. Evaluation methods for interpretable machine learning models require human perspective evaluation methods. The researcher measures human performance before and after the explanations of the model introduced. As explained by Murdoch et al. (2019)and Doshi-Velez & Kim (2017), the three interpretability evaluation methods are (1) Application-grounded evaluation by conducting human experiments in a real-world application, (2) Human-grounded evaluation is conducting simpler human-subject experiments that maintain the essence of the target application and (3) Functionally-grounded evaluation requires no human experiments and uses a formal definition of interpretability as a proxy for explanation quality. These evaluation methods can suitably be used to measure relevancy and descriptive accuracy.

V. Challenges in interpretability

Errors can be introduced in the Machine Learning interpretations when approximating (Murdoch et al., 2019) the underlying data relationships with a model (predictive accuracy) and descriptive accuracy is subject to misinterpretation based on the end-users understanding. The biases (Mehrabi et al., 2021) introduced by human experts during the data collection and processing can negatively influence the predictive and descriptive accuracy of the interpretable machine learning model. During the data collection, some data sets can be weighted more and others less by humans when they are equal. Facial recognition systems which don't have enough representative data sets for most of the human race can bias its decision based on race or ethnicity.

The options to select the machine learning model are either to choose simple easier to interpret and complex black box models which fit the data accurately. Model-based interpretation (Lipton, 2016) is best used when the relationship is simple whereas the post hoc analysis is used for complex models.

Due to the diversity of users' experience and expertise in the problem domain of different users of ML models, researchers can face difficulties to explain and users understanding the explanation.

According to the European Union (Goodman & Flaxman, 2016; Datta et al., 2016), the users have the right to an explanation of algorithms that affect their day-to-day activities. Additionally, laws and standards change over time which will affect interpretability in terms of how and why interpretability is introduced to end-users. The desire of users to fully understand the model may raise security concerns and privacy issues. The complexity (Murdoch et al., 2019) of modern deep neural networks may rise a major computational cost and difficulty of implementation issues in the interpretability. The undefined boundary of Interpretability may cause the end-user exhaustive new requirements.

Predictive accuracy (Lipton, 2016; Murdoch et al., 2019) of a model is simple to measure, but the descriptive accuracy and relevancy are difficult to quantify and can lead to imperfect interpretations. The post-hoc interpretations can potentially mislead to wrong conclusions. Post-hoc explanations are subject to different interpretation results based on the user. In model interpretability, reducing the number of parameters in introducing sparsity to the model requires understanding the data-specific organization and modeling which is difficult. Introducing simulatability is difficult when the model is complex.

Deep Neural Networks (Lipton, 2016), lack simulatability and decomposability because parameters in the hidden layer do not have an intuitive explanation, and is also difficult to know the computations of neural networks.

Predictive accuracy (Murdoch et al., 2019) is easy to measure whereas descriptive accuracy and relevancy are both difficult to evaluate and quantify.

VI. Thoughts and reflections

The researcher suggests and reflects some ideas and considerations to improve the interpretations of machine learning models.

Machine learning models (Mehrabi et al., 2021) are vulnerable to biases that deviate their decision from the expected predictive and descriptive accuracy results. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software in (Mehrabi et al., 2021), measures the risk of a person recommitting another crime. The judge uses the software to release an offender or stay in prison. A bias against African-Americans was found in the software that on producing higher false-positive rates (likely recommitting a crime) than others. The prediction result is biased. These biases arise from hidden or neglected biases in data and algorithms (Mehrabi et al., 2021). Biases can't be avoided but minimized by selecting the right machine learning model, and choosing correctly pre-processed training data sets. Following and monitoring Machine learning projects using standard data science life cycle (NICK HOTZ, 2021), the developers and researchers can minimize biases in machine learning models that negatively impact predictive accuracy (model interpretability) and descriptive accuracy (post-hoc interpretability). The researchers can choose the models that maximize predictive and descriptive accuracy. The predictive accuracy can create a tradeoff (i.e. if either predictive or descriptive accuracy is as good as expected). This tradeoff can be solved using the relevancy of the PDR Framework (Murdoch et al., 2019), in which users study or using the explanation in the real world will be used to evaluate how much it is important to the end-users.

Predictive accuracy (Murdoch et al., 2019) is easy to quantify. In the case of descriptive accuracy, it has a higher chance of misleading different decisions according to the users' understanding and experience. Descriptive accuracy and relevancy are both difficult to evaluate and quantify. Further research is demanding on how to quantify the interpretability of machine learning models. The domain stakeholders need to be engaged to establish a wider acceptance and make it compatible with real-world scenarios. Governments and organizations should have set and followed universal standards to ensure interpretability.

Researchers and developers should have to develop frameworks and standards for industries. The data privacy for the users should have to be ensured. Similar to the European Union (Goodman & Flaxman, 2016) law on algorithmic transparency and explanation, guidelines, laws, and other ethical considerations should have to be regulated on how machine learning models are going to interpret.

Developing interpretability frameworks similar to PDR (Murdoch et al., 2019) is relevant to have a common understanding of machine learning interpretability.

If the internal working of machine learning models is explained and exposed due to the nature of interpretability the machine learning models can be attacked by hackers.

Additionally, human laws and standards change over time which will affect interpretability in terms of how and why interpretability is introduced to end-users. With the fast pace of innovation of ML-powered technologies with different specifications, data usage scenarios and new solutions establishing new interpretability laws and standards are difficult. Further research and intergovernmental engagement are required to establish new laws and standards.

Generally, the area of machine learning interpretability demands extensive research at an equivalent pace with the rise of machine learning and artificial intelligence technologies.

VII. Conclusion

The notion of interpretability is important and difficult to define. Interpretability is the ability to give details or to present in comprehensible terms to a human. Governments and industries should have established interpretability in the implementation and use of machine learning models with the desiderata of interpretability. The most common interpretability of machine learning models are model-based and post hoc analysis.

References

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, 598–617. https://doi.org/10.1109/SP.2016.42

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. http://arxiv.org/abs/1702.08608

Goodman, B., & Flaxman, S. (2016). *European Union regulations on algorithmic decision-making and a "right to explanation."* https://doi.org/10.1609/aimag.v38i3.2741

Lipton, Z. C. (2016). *The Mythos of Model Interpretability*. http://arxiv.org/abs/1606.03490

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3457607

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Interpretable machine learning: definitions, methods, and applications*. https://doi.org/10.1073/pnas.1900654116

NICK HOTZ. (2021, February 28). *What is a Data Science Life Cycle?* . Https://Www.Datascience-Pm.Com/Data-Science-Life-Cycle/.