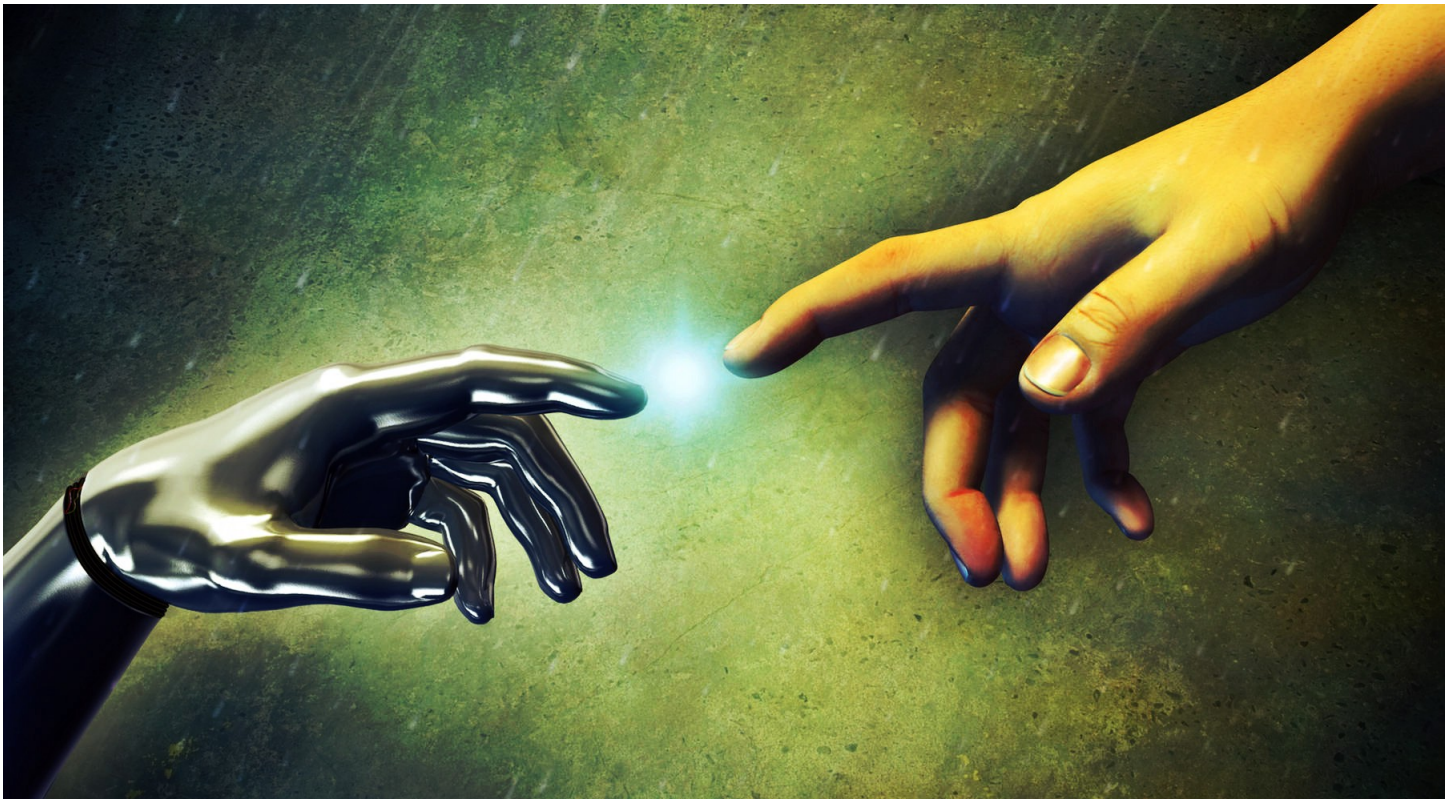
EXPLAINABLE ARTIFICIAL INTELLIGENCE (PART 1)

The Importance of Human Interpretable Machine Learning

A brief introduction into human interpretable machine learning and model interpretation

Dipanjan (DJ) Sarkar
May 25, 2018 · 13 min read



Introduction

This article is the first in my series of articles aimed at ***'Explainable Artificial Intelligence (XAI)'***. The field of Artificial Intelligence powered by Machine Learning and Deep Learning has gone through some phenomenal changes over the last decade. Starting off as just a pure academic and research-oriented domain, we have seen widespread industry adoption across diverse domains including retail, technology, healthcare, science and many more. Rather than just running lab experiments to

publish a research paper, the key objective of data science and machine learning in the 21st century has changed to tackling and solving real-world problems, automating complex tasks and making our life easier and better. More than often, the standard toolbox of machine learning, statistical or deep learning models remain the same. New models do come into existence like *Capsule Networks*, but industry adoption of the same usually takes several years. Hence, in the industry, the main focus of data science or machine learning is more ***'applied'*** rather than theoretical and effective application of these models on the right data to solve complex real-world problems is of paramount importance.

A machine learning model by itself consists of an algorithm which tries to learn latent patterns and relationships from data without hard-coding fixed rules. Hence, explaining how a model works to the business always poses its own set of challenges. There are some domains in the industry especially in the world of finance like insurance or banking where data scientists often end up having to use more traditional machine learning models (linear or tree-based). The reason being that model interpretability is very important for the business to explain each and every decision being taken by the model. However, this often leads to a sacrifice in performance. This is where complex models like ensembles and neural networks typically give us better and more accurate performance (since true relationships are rarely linear in nature). We, however, end up being unable to have proper interpretations for model decisions. To address and talk about these gaps, I will be writing a series of articles where we will explore some of these challenges in-depth about explainable artificial intelligence (XAI) and human interpretable machine learning.

Outline for this Series

Some of the major areas we will be covering in this series of articles include the following.

Part 1: The Importance of Human Interpretable Machine Learning

- Understanding Machine Learning Model Interpretation
- Importance of Machine Learning Model Interpretation
- Criteria for Model Interpretation Methods
- Scope of Model Interpretation

Part 2: Model Interpretation Strategies

- Traditional Techniques for Model Interpretation
- Challenges and Limitations of Traditional Techniques
- The Accuracy vs. Interpretability trade-off
- Model Interpretation Techniques

Part 3: Hands-on Model Interpretation — A comprehensive Guide

- Hands-on guides on using the latest state-of-the-art model interpretation frameworks
- Features, concepts and examples of using frameworks like ELI5, Skater and SHAP
- Explore concepts and see them in action — Feature importances, partial dependence plots, surrogate models, interpretation and explanations with LIME, SHAP values
- Hands-on Machine Learning Model Interpretation on a supervised learning example

Part 4: Hands-on Advanced Model Interpretation

- Hands-on Model Interpretation on Unstructured Datasets
- Advanced Model Interpretation on Deep Learning Models

This content will be covered across several articles in this series as highlighted above to keep things concise and interesting, so that everyone can get some key takeaways from every article.

Motivation

Working as a Data Scientist in the industry and mentoring people in the field, I have seen that data science is still often perceived as a black box capable of performing magic or alchemy to give people what they want. However, the harsh reality is that without a reasonable understanding of how machine learning models or the data science pipeline works, real-world projects rarely succeed. Considering any data science project in the real world, you will typically have a business aspect and the technical or solution aspect. Now, data scientists typically work to build models and provide solutions for the business. However, the business may not know the intricate details of how a model might work. But since this very model will be making a lot of decisions for

them in the end, they do have a right to pose the question, ***"How can I trust your model?"*** or ***"How does your model really make its decisions?"*** Answering these questions is something data science practitioners and researchers have been trying over several years now.

Data Science practitioners will know that there exists a typical ***model interpretability vs. model performance trade-off***. A point to remember here is that model performance is not the run-time or execution performance, but how accurate the model can be in making decisions. There are several models including simple linear models or even tree-based models, which can easily explain the decisions taken by the model to arrive at a particular insight or prediction, but you might need to sacrifice model performance since they always do not yield the best results due to inherent problems of high bias (linear models) or high variance, leading to overfitting (fully grown tree models). More complex models like ensemble models and the more recent deep learning family of models often yield better performance, but are perceived as black-box models, because it is extremely difficult to explain how the model might be really making its decisions under the hood.

While some might argue that if something is working well (for the time being), why question how it works? However, being humans, logic and reasoning is something we

adhere to for most of our decisions. Hence, the paradigm shift towards artificial intelligence (AI) making decisions will no doubt be questioned. There are a lot of real-world scenarios where biased models might have really adverse effects. This includes *predicting potential criminals, judicial sentencing risk scores, credit scoring, fraud detection, health assessment, loan lending, self-driving* and many more where model understanding and interpretation is of utmost importance. The same is highlighted by renowned Data Scientist and Author Cathy O' Neil in her acclaimed book, ***'Weapons of Math Destruction'***.

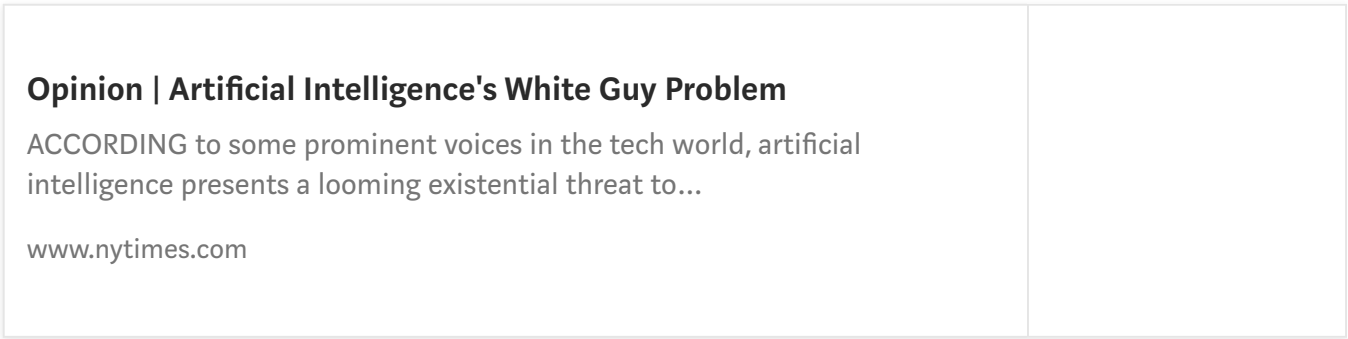
Main

Weapons of Math Destruction has been Longlisted for the National Book Award! Book description: A former Wall Street...

weaponsofmathdestructionbook.com

Renowned researcher and author, Kate Crawford, talked about these very aspects of the implications of bias in machine learning and it's effects on society in the NIPS 2017 Keynote, ***'The Trouble with Bias'***.

Interested readers should also definitely check out her famous article on the NY Times, ***'Artificial Intelligence's White Guy Problem'*** where she shows us examples of machine learning applications including image categorization, criminal risk predictions, delivery service availability and many more were biased and yielded unfavorable outcomes for the black community. All these real-world scenarios are implications of how important model interpretation should be, and if we want to leverage machine learning to solve these problems.



In the past year, I have seen the need for model interpretation while solving problems in the industry and also when I was writing my recent book ***'Practical Machine Learning with Python'***. During this time, I have had the chance to interact with the wonderful folks at *DataScience.com* who are very much aware of the need and importance of human interpretability in machine learning models. They have been actively working on a solution and have open-sourced the popular python framework, ***Skater***. We will be taking a deep-dive into Skater and also do some hands-on model interpretation in this series of articles. Besides this we will also do a comprehensive coverage of other model interpretation frameworks like ***ELI5*** and ***SHAP***!

Understanding Machine Learning Model Interpretation

Machine Learning has seen widespread industry adoption only in the last couple of years. Hence, model interpretation as a concept is still mostly theoretical and subjective.

> *Any machine learning model at its heart has a response function which tries to map and explain relationships and patterns between the independent (input) variables and the dependent (target or response) variable(s).*

When a model predicts or finds our insights, it takes certain decisions and choices. Model interpretation tries to understand and explain these decisions taken by the response function i.e., the what, why and how. The key to model interpretation is transparency, the ability to question, and the ease of understanding model decisions by

humans. The three most important aspects of model interpretation are explained as follows.

1. **What drives model predictions?** We should have the ability to query our model and find out latent feature interactions to get an idea of which features might be important in the decision-making policies of the model. This ensures ***fairness*** of the model.

2. **Why did the model take a certain decision?** We should also be able to validate and justify why certain key features were responsible in driving certain decisions taken by a model during predictions. This ensures ***accountability*** and reliability of the model.

3. **How can we trust model predictions?** We should be able to evaluate and validate any data point and how a model takes decisions on it. This should be demonstrable and easy to understand for key stakeholders that the model works as expected. This ensures ***transparency*** of the model.

> ***Interpretability*** *also popularly known as* ***human-interpretable interpretations (HII)*** *of a machine learning model is the extent to which a human (including non-experts in machine learning) can understand the choices taken by models in their decision-making process (the how, why and what).*

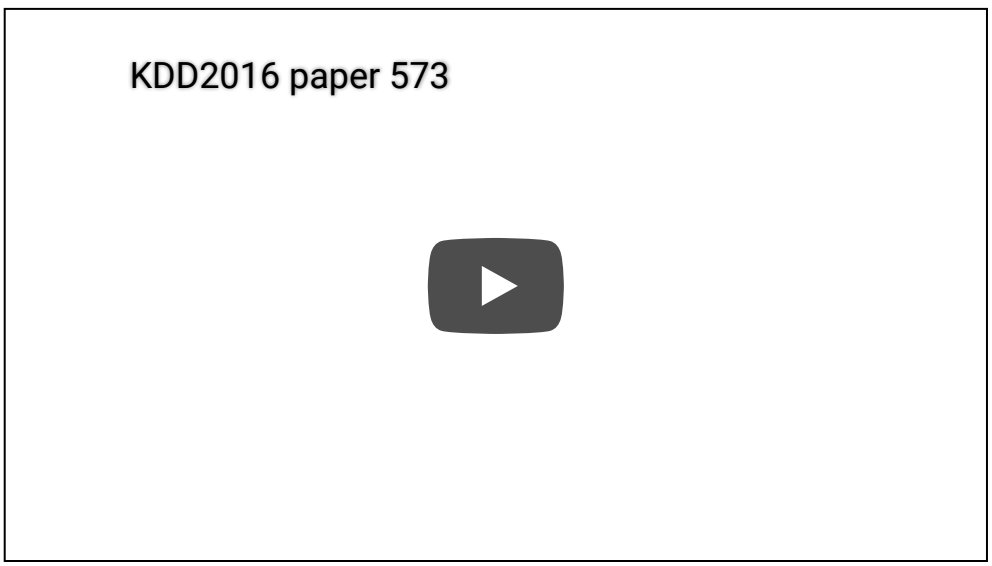
When comparing models, besides model performance, a model is said to have a better interpretability than another model if its decisions are easier to understand by a human than the decisions from the other model.

The Importance of Machine Learning Model Interpretation

When tackling machine learning problems, data scientists often have a tendency to fixate on model performance metrics like accuracy, precision and recall and so on (This is important no doubt!). This is also prevalent in most online competitions around data science and machine learning. However, metrics only tell a part of the story of a model's predictive decisions. Over time, the performance might change due to model concept drift caused by various factors in the environment. Hence, it is of paramount importance to understand what drives a model to take certain decisions.

Some of us might argue if a model is working great why bother digging deeper? Always remember that when solving data science problems in the real-world, for the business to trust your model predictions and decisions, they will keep asking the question, ***"Why should I trust your model?"*** and this makes perfect sense. Would you be satisfied with a model just predicting and taking decisions (the ***what***) like if a person has cancer or diabetes, if a person might be a risk to society or even if a customer will churn? Maybe not, we might prefer it more if we could know more about the model's decision process (the ***why*** and ***how***). This gives us more transparency into why the model makes certain decisions, what might go wrong in certain scenarios and over time it helps us build a certain amount of trust on these machine learning models.

The key takeaway from this section is that it is high time we stop seeing machine learning models as black boxes and try and analyze not just data, but how models make decisions. In-fact, some of the key steps towards this path was started by the famous paper, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* by M. T. Ribeiro, S. Singh and C. Guestrin, SIGKDD 2016, where they introduce the concept of LIME (Local Interpretable Model-Agnostic Explanations) which we will be covering in the next article in detail.



They mention some key points in their paper which are worth remembering.

> *Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when*

> *choosing whether to deploy a new model.*
>
> *Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: if the users do not trust a model or a prediction, they will not use it.*

Interested people should also check out their talk in the KDD conference around their paper in model interpretation.

This is something we have discussed several times in this article and is one of the key differentiators which determines the success of data science projects in the industry. This drives the urgency around the need and importance of model interpretation.

Criteria for Machine Learning Model Interpretation Methods

There are specific criteria which can be used for categorizing model interpretation methods. An excellent guide to this is mentioned in *"Interpretable Machine Learning, A Guide for Making Black Box Models Explainable"* by Christoph Molnar, 2018

- **Intrinsic or post hoc?** Intrinsic interpretability is all about leveraging a machine learning model which is intrinsically interpretable in nature (like linear models, parametric models or tree based models). Post hoc interpretability means selecting and training a black box model (ensemble methods or neural networks) and

applying interpretability methods after the training (feature importance, partial dependency plots). We will focus more on post hoc model interpretable methods in our series of articles.

- **Model-specific or model-agnostic?** Model-specific interpretation tools are very specific to intrinsic model interpretation methods which depend purely on the capabilities and features on a per-model basis. This can be coefficients, p-values, AIC scores pertaining to a regression model, rules from a decision tree and so on. Model-agnostic tools are more relevant to post hoc methods and can be used on any machine learning model. These agnostic methods usually operate by analyzing (and perturbations of inputs) feature input and output pairs. By definition, these methods do not have access to any model internals like weights, constraints or assumptions.

- **Local or global?** This classification of interpretation talks about if the interpretation method explains a single prediction or the entire model behavior? Or if the scope is somewhere in between? We will talk more about global and local interpretations soon.

This is not an exhaustive set of criteria for classifying interpretable methods since this is still an emerging field, but this can be a good yardstick to compare and contrast across multiple methods.

Scope of Machine Learning Model Interpretation

How do we define the scope and boundaries of interpretability? Some useful aspects can be the transparency, fairness and accountability of a model. *Global* and *Local* model interpretations are clear ways to define the scope of model interpretation.

Summarizing Global and Local Interpretation (Source: DataScience.com)

Global Interpretations

This is all about trying to understand ***"How does the model make predictions?"*** and ***"How do subsets of the model influence model decisions?".*** To comprehend and interpret the whole model at once, we need global interpretability. Global interpretability is all about being able to explain and understand model decisions based on conditional interactions between the dependent (response) variable(s) and the independent (predictor) features on the complete dataset. Trying to understand feature interactions and importances is always a good step towards understanding global interpretation. Of course, visualizing features after more than two or three dimensions becomes quite difficult when trying to analyze interactions. Hence, often looking at modular parts and subsets of features, which might influence model predictions on a global knowledge, helps. Complete knowledge of the model structure, assumptions and constraints are needed for a global interpretation.

Local Interpretations

This is all about trying to understand ***"Why did the model make specific decisions for a single instance?"*** and ***"Why did the model make specific decisions for a group of instances?"***. For local interpretability, we do not care about the inherent structure or assumptions of a model and we treat it as a black box. For understanding prediction decisions for a single datapoint, we focus specifically on that datapoint and look at a local subregion in our feature space around that point, and try to understand model decisions for that point based on this local region. Local data distributions and feature spaces might behave completely different and give more accurate explanations as opposed to global interpretations. The Local Interpretable Model-Agnostic Explanation (LIME) framework is an excellent method which can be used for model-agnostic local interpretation. We can use a combination of global and local interpretations to explain model decisions for a group of instances.

Model Transparency

This is all about trying to understand ***"How was a model created from algorithms and features?".*** We know that typically a machine learning model is all about leveraging an algorithm on top of data features to build a representation which maps inputs to potential outputs(responses). Transparency of a model can be trying to understand more technical details of how models are built and what might influence its decisions.

This can be weights of a neural network, weights of a CNN filter, linear model coefficients, the nodes and splits of a decision tree. However, since the business may not be very well-versed in these technical details, trying to explain model decisions with agnostic local and global interpretation methods helps in showcasing model transparency.

Conclusion

Model Interpretation is something which can make or break a real-world machine learning project in the industry and helps us come one step closer to explainable artificial intelligence (XAI). Let's try and work towards human-interpretable machine learning and XAI to demystify machine learning for everyone and help increase the trust in model decisions.

. . .

What's next?

In Part 2 of this series, we will be covering the following aspects of explainable artificial intelligence with regard to machine learning model interpretation.

- Traditional Techniques for Model Interpretation
- Challenges and Limitations of Traditional Techniques
- The Accuracy vs. Interpretability trade-off
- Model Interpretation Techniques

. . .

Thanks to ***Matthew Mayo*** for editing and featuring this article on ***KDNuggets***.

Thanks to all the wonderful folks at DataScience.com and especially *Pramit Choudhary* for building an amazing model interpretation framework, ***Skater,*** and helping me out with some excellent content for this series.

I cover a lot of examples of machine learning model interpretation in my book, ***"Practical Machine Learning with Python"***. The code is open-sourced for your

benefit!

If you have any feedback, comments or interesting insights to share about my article or data science in general, feel free to reach out to me on my LinkedIn social media channel.

Dipanjan Sarkar - Data Scientist - Intel Corporation | LinkedIn

View Dipanjan Sarkar's profile on LinkedIn, the world's largest professional community. Dipanjan has 5 jobs listed on...

www.linkedin.com