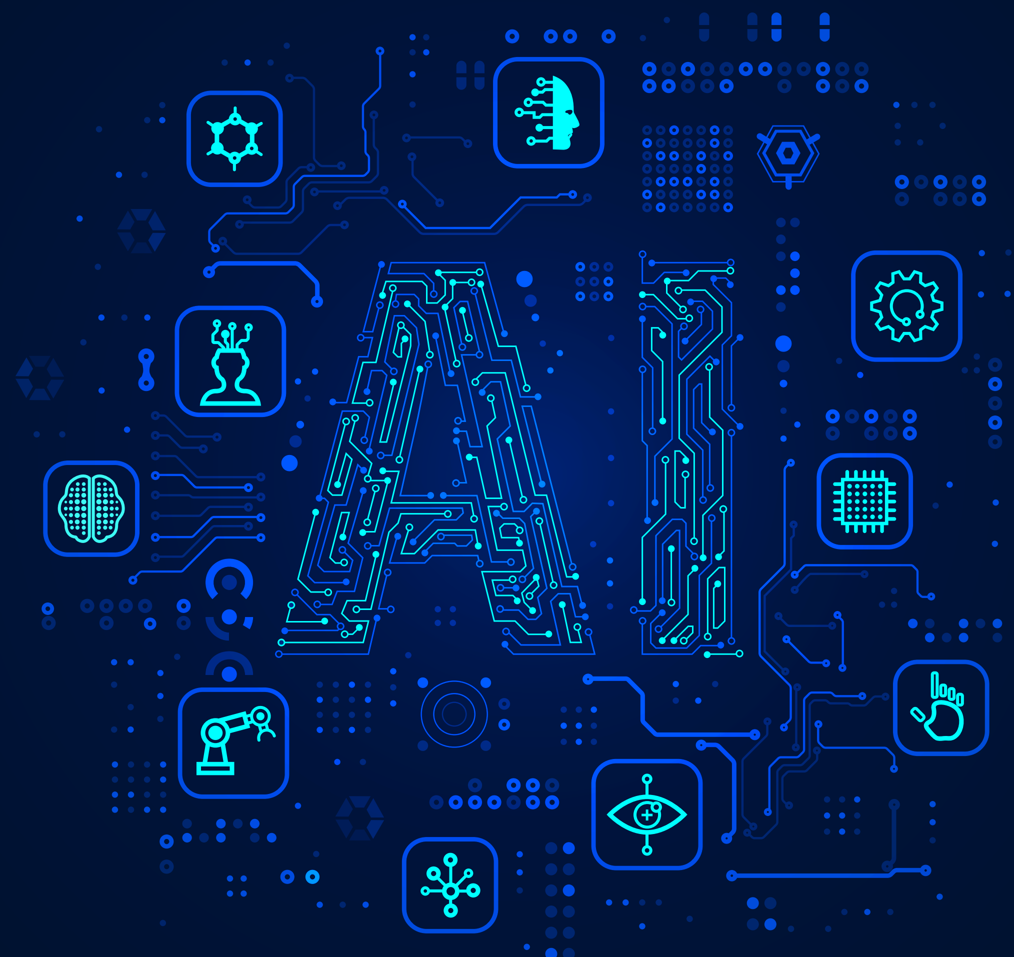
LTI

Let's Solve

A Larsen & Toubro Group Company



Whitepaper

3 X's of Explainable AI

Author: **Arup Duttaroy,** Dy. Head- Mosaic, LTI

Table of Contents

1 Introduction

Machine Learning & Artificial Intelligence have provided some significant breakthroughs in diverse fields in recent years. Areas like financial services, healthcare, retail, transportation, and more have been using machine learning systems in one way or another, and the results have been promising. Artificial intelligence and machine learning are now technology interventions that almost any company can use to find practical answers to difficult business problems. The sorts of decisions and predictions being made by AI-enabled systems is becoming much more profound, and in many cases, critical to life, death, and personal wellbeing. As artificial intelligence becomes an increasing part of our daily lives, from the image and facial recognition systems to machine learning-powered predictive analytics, conversational applications, autonomous machines, and hyper-personalized systems, the need to trust these AI-based systems is becoming of paramount importance.

2 The 3 X's of Explainable AI

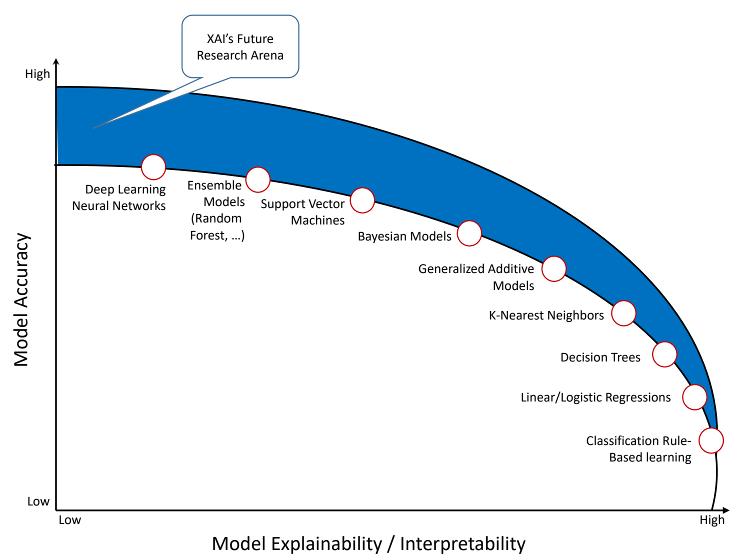
As we all know, existence of AI/ML, at the end of the day, are to aid in decision making. When humans make decisions, they can explain their thought process behind it, whether or not the rationale is driven by observation, intuition, experience or logical thinking ability. Basic ML algorithms like decision trees can be explained by following the tree path, which led to the decision. The deep layers of neural networks have a magical ability to recreate the human mind and its functionalities, but such deep layers of complex AI algorithms are often incomprehensible by human intuition and are quite opaque. The AI/ML process essentially follows a "black-box" concept and is faced with a typical technical challenge known as "Interpretability Problem" because of which even the Data Scientists, who designed such AI/ML algorithms in the first place, cannot explain why the AI algorithm arrived at a specific decision or predicted a specific outcome. The laymen end-user may not simply trust the machine's predictions without contextual proof and reasoning. Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. This article delves on three dimensions of XAI that everyone should be aware of, hitherto referred as the 3X's of Explainable AI. They are:

1. X for Explainability vs. Accuracy
2. X for Explainability by Design, whose scope is to explain the entire AI/ML algorithm itself
3. X for Explainability with Local Interpretation, whose scope is to explain an individual prediction made by the AI/ML algorithm in question.

2.1 The 1st X – Explainability v/s accuracy tradeoff

Explainability is a multifaceted topic. It encompasses both individual models and the larger systems that incorporate them. It refers not only to whether the decisions a model generates are interpretable, but also whether or not the whole process and intention surrounding the model can be properly accounted for. Data Scientists try to have an efficient trade-off between Accuracy and Explainability, along with a great human-computer interface, which can help translate the model to understandable representation for the end users.

When we build an AI/ML model, whether Supervised models like regression or classification and/or Unsupervised models like Association or Clustering, we have to train the model on sample data referred as "training dataset" and also validate the model with yet another sample data referred as the "test dataset". We have to resort to iterative model development approach in order to maximize the performance of the model. Model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training data. The better a model can generalize to 'unseen' data, the better predictions and insights it can produce, which in turn, deliver more business value. When we have to evaluate a model, we do consider accuracy but what we majorly focus on is how much robust our model is, how will it perform on a different dataset and how much flexibility it has to offer. A model is considered to be robust when it has realized and learned about the data correctly and desirably, and are able to make predictions that are close to the actual values. One must have experienced a situation where the model achieves good accuracy on one test set, but fails to perform equally when new data is provided (of the same nature). As one starts tweaking the model to handle such situation, the model starts becoming more complex and difficult to explain. By their very nature, some models are more explainable than others. In fact, if we try to map the various kinds of model into an accuracy continuum in one dimension and explainability continuum in another, we get an interesting insight as shown in the figure below:



< BACK

We see that explainable models are easily understandable, but their accuracy is not as good, as they are simple. Accurate models, on the other hand, work well, but aren't as explainable, as they are complicated. Now a question which looms large is, "Do we need to 'lower down' the accuracy of AI algorithms to make them explainable? The main issue with explainable AI is whether it can accurately fulfill the task it was designed for. The trade-off decision to be made should depend on the field of application of the algorithm and the end-user base who are likely to be impacted i.e. to whom it is accountable.

When dealing with technical users who are acquainted with the field and have high trust level, especially tech companies or researchers, having highly accurate models would be preferred over high explainability as performance is very important. But while dealing with laymen users it would be difficult to garner their trust. The machine learning approach is very different specially in banks, insurance companies, healthcare providers and other regulated industries. The reason is mainly that they are prone to **legal or even ethical requirements,** which tend to limit more and more the use of accurate black box models. Such institutions are answerable for their decisions and processes. In such cases, simple explainable albeit less efficient algorithms would do.

Hence the performance vs. explainability trade-off should be decided according to the application domain and the target users concerned. If one wants the system to be explainable, they need to make do with a simpler system that isn't as powerful or accurate. Simple systems can give a prediction to the user, but the ultimate action should be taken by the user. When performance matters most, even with a complicated model, one should opt for the trust that comes from making sure that one is able to verify that the system does, in fact, work.

For Data Scientists, XAI is a big hope to usher in the next-gen AI systems, which can understand the context in which they function and can characterize the real-world phenomena, with the main aim to:

a. Explain the decisions and its process (why a prediction & why not)

b. Understand its strengths and weaknesses (success & failure and reasons thereof)

c. Convey how the system may behave in the future (drift vs. re-fitment)

d. Offer insight on how to correct the mistakes (when it erred and why)

The final goal of Explainable AI (XAI) is to have a toolkit library of AI/ML algorithms and Human-Computer-Interface modules, which can help understand or interpret AI implementations. That is the focus of the next two X's, which deals with the scope of explainability or interpretability of AI algorithms and their associated toolkits and/or methods.

2.2 The 2nd X – Explainability by design is vital

The terms Explainability and Interpretability are such that they can be used interchangeably. By definition, interpretability is the degree to which a human can understand the cause of a decision and can consistently predict the model's result.

As we saw above, generally, linear models, as well as tree-based models, can be easily interpreted because of their intuitive way of getting to the predictions, but we may need to sacrifice on accuracy since these models are simple and can easily under- or overfit depending on the problem. On the other hand, we have more complex models like ensemble models (e.g. Random Forest, XGBoost, ...) as well as deep neural networks which are especially hard to interpret because of their complexity. For real-world problems like fraud detection, self-driving cars or loan lending the model doesn't only need to perform well, but also needs to be easily interpretable so we can see why a loan was/wasn't approved and use our domain-expertise to validate or correct the decision.

As mentioned above, model interpretability tries to understand and explain the steps and decision a machine learning model takes when making predictions. It gives us the ability to question the model's decision and learn about the following aspects:

- **What features/attributes are important to the model?** We should be able to extract information about what features are important, as well as how features interact to create powerful information.
- **Why did the model come to this conclusion?** We should also have the ability to extract information about specific predictions in order to validate and justify why the model produced a certain result.

As we saw earlier, AI/ML process works in two phases: a build and training phase using a training dataset, which is a sample of real-world observations and then the usage phase when a trained model is deployed to predict individual real-world events in real life. Similarly, the explainability or interpretability of models also has two separate scope: whether the interpretation method explain an individual prediction or the entire model behavior? The first kind is known as "Local Interpretation" and is the topic for our third X, while the second one is referred as "Global Interpretation", which is the subject of this second X. Global Interpretation is also referred as "Explainability by Design".

An algorithm trains a model that produces the predictions. Each step can be evaluated in terms of transparency or interpretability. Questions that can arise are:

a. **How does the algorithm create the model?** This question is about Algorithm Transparency, which is about how the algorithm learns a model from the data and what kind of relationships it can learn. If one uses convolutional neural networks to classify images, one can explain that the algorithm learns edge detectors and filters on the lowest layers. This is an understanding of how the algorithm works, but not for the specific model that is learned in the end, and not for how individual predictions are made. Algorithm transparency only requires knowledge of the algorithm and not of the data or the learned model.

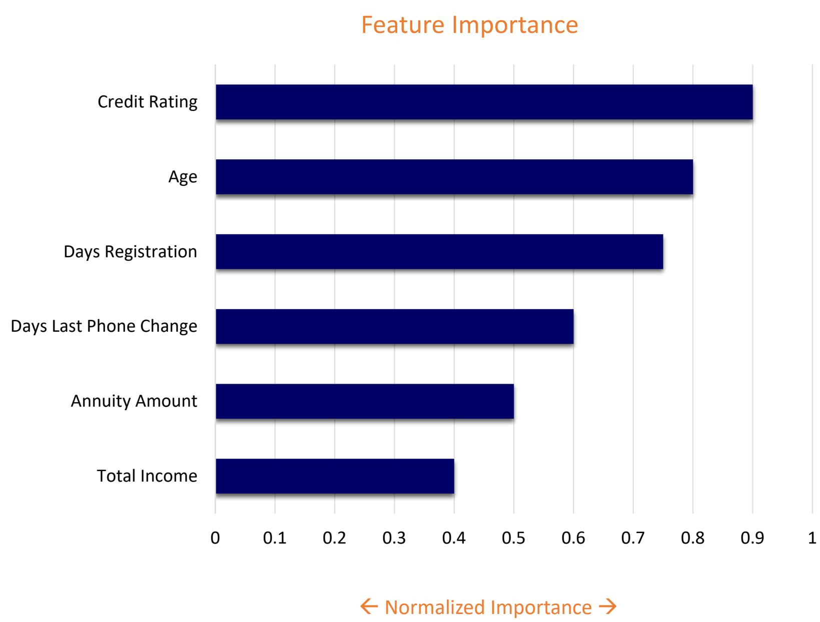
b. **How does the trained model make predictions?** This question, on the other hand, is about comprehending the entire model at once i.e. global interpretability. To explain the global model output, we need the trained model, knowledge of the algorithm and the data. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures.

c. **Which features are important and what kind of interactions between them take place?** Global model interpretability helps to understand the distribution of the target outcome (dependent variable) based on the features or the independent variables.

d. **How do parts of the model affect predictions?** This question is again related to global interpretation but on a modular level. While global model interpretability may be out of reach, there is a good chance of understanding, at least for some models, on a modular level. Not all models are interpretable at a parameter level.

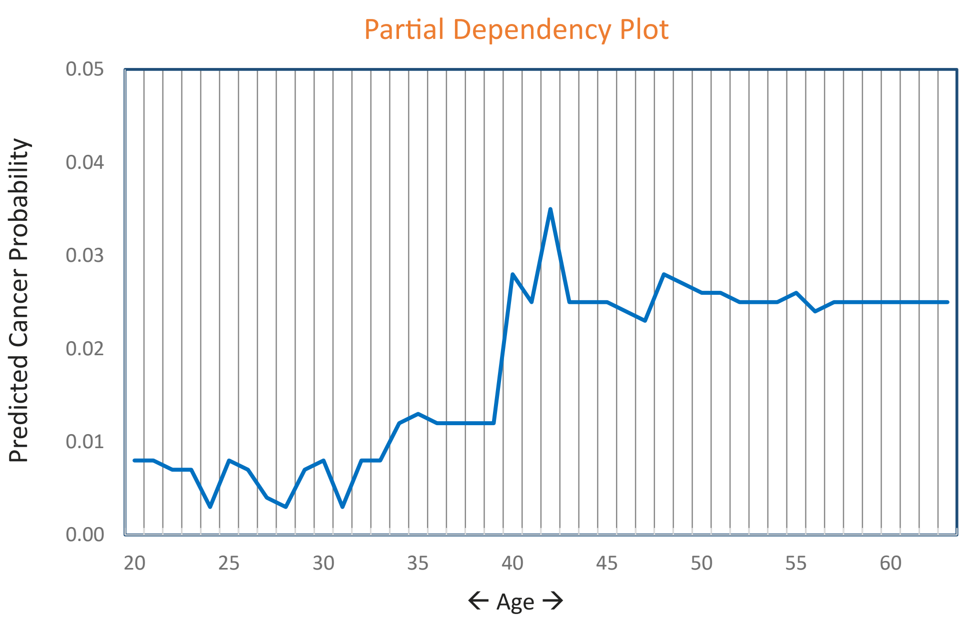
While, global model interpretability is very difficult to achieve in practice, it is absolutely vital for explainability because it tries to explain what features drive predictions and what features are completely useless for the cause right at the design stage itself. That is precisely why it is referred as "Explainability-by-Design". Using this knowledge, we can make decisions about the data collecting process, create dashboards to explain our model or use our domain knowledge to fix obvious bugs. Most global interpretation methods work by investigating the conditional interactions between the dependent variable and the independent variables (features) on the complete dataset. They also create and use extensive visualizations which are mostly easy to understand but contain a huge amount of useful information for analyzing our model. The two most used global model interpretation techniques are:

Feature importance - We can use feature importance to get an understanding of how important a model thinks a feature is for making predictions. The concept of feature importance is really straightforward. Using feature importance, we measure the importance of a feature by calculating the increase in the error of a given model after permuting/shuffling the feature values of a given feature. A feature is "important" if permuting it increases the model error. This is because in that case, the model relied heavily on this feature for making right prediction. On the other hand, a feature is "unimportant" if permuting it doesn't affect the error by much or doesn't change it at all. Feature importance is a very popular because it is a simple technique that gives us highly compressed global insights about the importance of a feature. Also, it does not require retraining the model which is always an advantage because of the saving of computing time. However, it has some disadvantages as well. For instance, it isn't clear whether we should use the training or testing set for calculating the feature importance.



Furthermore, because of the permutation process results can vary heavily when repeating the calculation. A typical feature importance plot is shown in the figure on the right, which is related to a problem to decide whether to sanction a loan to a person or not. Another problem is that correlation between features can bias feature importance by producing unrealistic instances or by splitting the importance between the two correlated features. Most libraries like Scikit-Learn, XGBoost as well as other machine learning libraries have their own feature importance methods. but if we want to get exact results when working with models from multiple libraries it is advantageous to use the same method to calculate the feature importance for every model, which is possible by using libraries like ELI5. ELI5 library allows users to visualize and debug various Machine Learning Models. It offers way more than just feature importance including library-specific features as well as a text-explainer.

- **Partial dependence plots** - Partial dependence plots (short PDP or PD plots) help us understand how a specific feature value effects prediction. This is extremely useful because it allows us to get interesting insight about specific feature values which can then be further analyzed or shared. The partial dependence can be calculated easily for categorical features. We get an estimate for each category by forcing all data instances to have the same category. For example, if we are interested in how gender affects the chance of having a heart disease, we can first replace all values of the gender column with the value male and average the predictions and then to the same using female as the value. A typical PDP is shown in the figure on the right, which is trying to show the dependency of age to predict the probability that a person can get diagnosed with the cancer disease. Calculating partial dependence for regression is way harder. For creating partial dependence plots, we can use the PDPbox library, which provides us with a few different well-designed plots including partial dependence plots for a single feature, as well as partial dependence plots for multiple features.



When it comes to explainability of Neural Networks, which are inherently difficult to explain, two other Ante-hoc methods (techniques that entail baking explainability into a model from the beginning i.e. from the design stage itself) which is gaining importance are:

- **Reversed Time Attention Model (RETAIN):** Researchers at Georgia-Tech developed the RETAIN model to help doctors understand the AI software's predictions. The patients' hospital visits data were sent to two RNN's (Recurrent Neural Network) both of which had attention mechanism. The attention mechanism helped explain which part the neural network was focusing on and which features helped influence its choice.

- **Bayesian Deep Learning (BDL):** BDL enables one to gauge how uncertain a neural network is about its predictions. These deep architectures can model complex tasks by leveraging the hierarchical representation power of deep learning, while also being able to infer complex multi-modal posterior distributions. Bayesian deep learning models typically form uncertainty estimates by either placing distributions over model weights, or by learning a direct mapping to probabilistic outputs. By knowing the weight distributions of various predictions and classes, we can tell a lot about what feature led to what decisions and the relative important of it.

Now that we are aware about Global Interpretations or better known as "Explainability by Design" and various methods for the same, it is time to look at Local Interpretations and its methods, which is the focus of the third & last X.

2.3 The 3rd X – Explainability with local interpretation is a must

When a trained model is deployed into production, it enters into the usage phase and starts predicting on real-world events or cases. For example, approving the loan application of one individual customer of a bank or trying to diagnose one particular patient whether he or she has Cancer or not. In such cases it is interesting to know Why did the model make a certain prediction for an instance or for a small group of instances?

We can zoom in on a single instance and examine what the model predicted for the particular input in question and explain why. If we look at an individual prediction, the behavior of the otherwise complex model might behave more pleasantly. Locally, the prediction might only depend linearly or monotonically on some features, rather than having a complex dependence on them. Local explanations can therefore be more accurate than global explanations. The methods that can make individual predictions more interpretable are mostly model-agnostic methods i.e. those that does not depend on the underlying algorithm being used. They are also post-hoc in nature i.e. methods that interpret a black box model, like a neural network or ensemble, by applying model interpretability methods after the fact i.e. after the training of the model has been completed. Post-hoc explainability method targets models that are not readily interpretable by design by resorting to diverse means to enhance their interpretability by using of most common ways humans explain systems and processes by themselves. Some of the popular model agnostic and post-hoc methods are as follows:

- **Text explanations** - This method deal with the problem of bringing explainability for a model by means of learning to generate text explanations that help explaining the results from the model using Natural Language Generation (NLG) techniques. Text explanations also include every method generating symbols that represent the functioning of the model. These symbols may portrait the rationale of the algorithm by means of a semantic mapping from model to symbols. An example will make it clear:

Considering the Model depicted below, Text explanation can look like:

Xi ⟶ Mφ ⟶ Yi

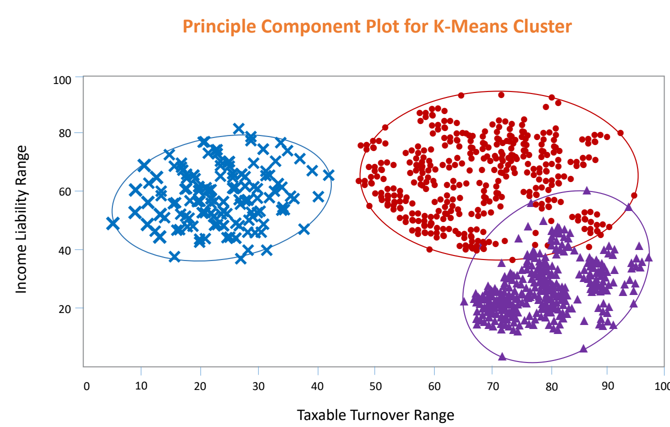
"The output y when x=7 is 50 because y = (x2 + 1)"

Text explanations are simple to understand but can also be applied for simple models like Linear or Polynomial regression. It can become unwieldy, if the number of parameters increases.

- **Visual explanations** - These techniques for post-hoc explainability aim at visualizing the model's behavior. Many of the visualization methods come along with dimensionality reduction techniques that allow for a human interpretable simple visualization. Visualizations may be coupled with other techniques, like text explanations, to improve their understanding, and are considered as the most suitable way to introduce complex interactions within the variables involved in the model to users not acquainted to ML modeling. Languages like Python, R, SAS etc. have many visualization libraries like matplotlib, ggplot, bokeh etc. using which such visualizations can be generated. A sample visualization is shown below for easy reference.

Considering the Model depicted below is a Clustering model, Visual explanation can look like:

Xi ⟶ Mφ ⟶ {Yi,Zi,Vi}

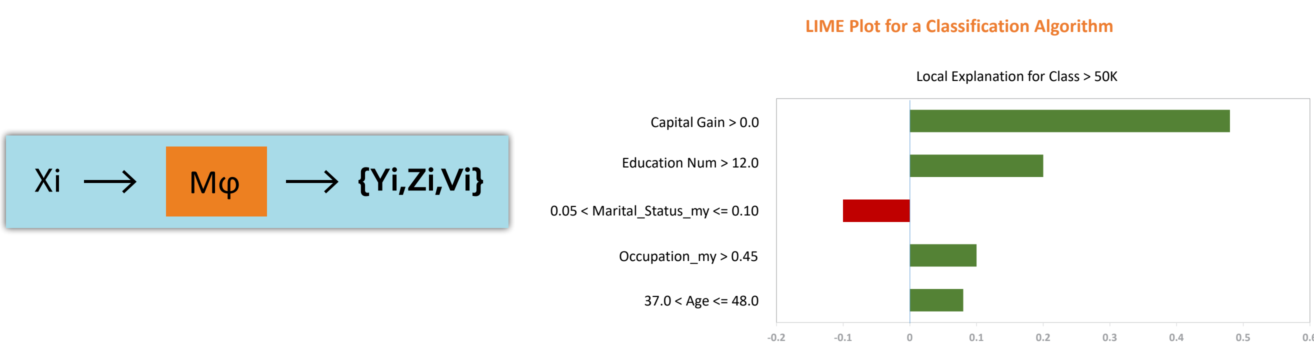


These two components explain 78.25% of the point variability

Visual explanations are simple to understand provided it is based on 2/3 variables at the max. It becomes unwieldy, if the number of parameters increases. Visual explanations can be applied to shallow ML models like Linear/Polynomial/Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), K-Means etc. It can also be applied for multi-layer neural networks, Convolutional Neural Networks (CNN) algorithms.

- **Local explanations** – This method tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model. These explanations can be formed by means of techniques with the differentiating property that these only explain part of the whole system's functioning. Among the most known approaches for local explanations is the technique of Local Interpretable Model-Agnostic Explanations (LIME) and all its variations. LIME is a local model interpretation technique using local surrogate interpretable models like Linear Regression or a Decision Tree, to approximate & explain the predictions of the underlying black-box model. LIME trains a surrogate model by generating a new dataset out of the datapoint of interest. Currently LIME supports tabular, text and image data. In the case of tabular data, LIME creates new samples by permuting each feature individually and for text and image data LIME generates the dataset by randomly turning single words or pixels on or off. LIME is best represented by a visual bar graph plot as shown below:

Considering the Model depicted below is a Classification model, local explanation using LIME can look like the graph on the right below:



LIME is a good pick for analyzing predictions because it can be used for any black box model no matter if it is a deep neural network or a Support Vector Machine (SVM). Furthermore, LIME is one of the only interpretation techniques that works for tabular, text and image data.

- **Explanations by example** – This method consider the extraction of data examples that relate to the result generated by a certain model, enabling to get a better understanding of the model itself. Similar to how humans behave when attempting to explain a given process, explanations by example are mainly centered in extracting representative examples that grasp the inner relationships and correlations found by the model being analyzed.

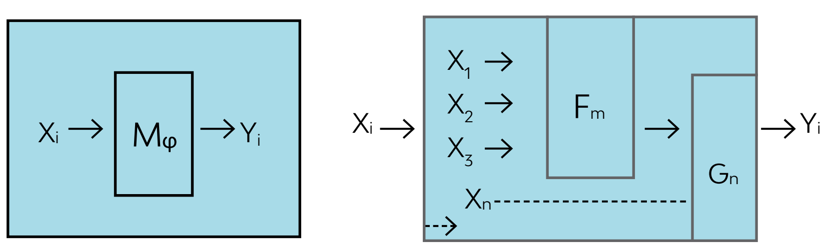
Considering the Model depicted below, Explanations by example can look like:

Explanatory examples of the model because $y = (x2 + 7)$:

1. "When x=0, then y = 7"
2. "When x=7, then y = 56"
3. "When x=20, then y = 407"

- **Explanations by simplification** – It collectively denote those techniques in which a whole new system is rebuilt based on the trained model to be explained. This new, simplified model usually attempts at optimizing its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score. Almost all techniques taking this path for model simplification are based on rule extraction techniques as illustrated below. An interesting byproduct of this family of post-hoc techniques is that the simplified model is, in general, easier to be implemented due to its reduced complexity with respect to the model it represents.

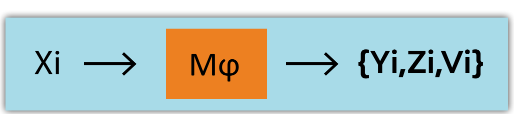
Considering the Model depicted below, Explanations by simplification can look like:

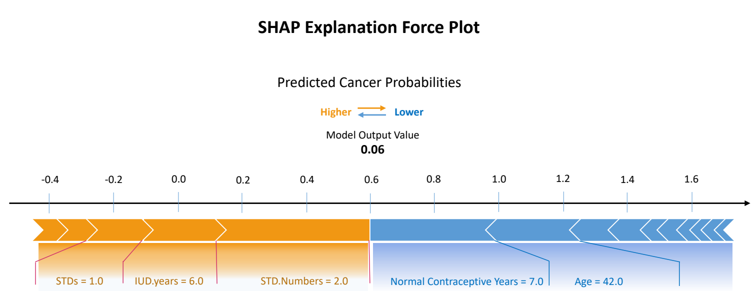


Explanations by simplification, are usually used for both Shallow ML models as well as Deep Learning models. It can be used for explaining Ensemble models (like Random Forrest, XGBoost etc.), Support Vector Machines (SVMs), as well as for explaining Multi-layer Neural Networks.

- **Feature relevance explanations techniques** – This method for post-hoc explainability clarify the inner functioning of an opaque model by computing a relevance score for its managed variables based on ranking or measuring the influence, relevance or importance each feature has in the prediction output by the model to be explained. These scores quantify the affection (sensitivity) a feature has upon the output of the model. A comparison of the scores among different variables unveils the importance granted by the model to each of such variables when producing its output. Feature relevance methods can be thought to be an indirect method to explain a model. One popular method of this technique is called SHAP (SHapley Additive exPlanations). This is a method to calculate an additive feature importance score for each particular prediction with a set of desirable properties (local accuracy, missingness and consistency) that its antecedents lacked. The Shapley value is a method for assigning payouts to players depending on their contribution to the total payout. In the case of machine learning the "game" is the prediction task for a data point. The "gain" is the prediction minus the average prediction of all instances and the "players" are the feature values of the data point. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. It is an excellent way of interpreting an individual prediction because it doesn't only give us the contribution of each feature value, but the scale of these contributions is also right, which isn't the case for other techniques like LIME. One can visualize feature attributions such as Shapley values as "forces". Each feature value is a force that either increases or decreases the prediction. The prediction starts from the baseline. The baseline for Shapley values is the average of all predictions. In the plot, each Shapley value is an arrow that pushes to increase (positive value) or decrease (negative value) the prediction. These forces balance each other out at the actual prediction of the data instance. it is best represented by a visual called SHAP explanation force plots as shown below:

Considering the Model depicted below is a Classification model, feature relevance explanation using SHAP Values can look like the graph on the right below:





< BACK

When using Shapley values the prediction and the average prediction is fairly distributed among the feature values of the instance. This makes Shapley values perfect for situations that need an exact explanation. Furthermore, Shapley value is the only local explanation method with solid theory behind it. It has some disadvantages as well. The Shapley value requires a lot of computing time. Most often, only the approximate solution is feasible. Another disadvantage is that we need access to the data to calculate the Shapley value. This is because we need the data to replace parts of the instance of interest. This problem can only be avoided if we can create data instances that look like real data instances.

For XAI, both Global Interpretation and Local Interpretation are equally important techniques. Good thing is there are library's available, open source or proprietary, to cater to these. A good platform for the Data Scientists, will offer this kind of capabilities and user interfaces out-of-the-box to make the lives of the Data Scientists better.

3 Benefits of Explainable AI

Every day, we see how AI can help the mankind and make a positive difference in our lives—from helping radiologists detect lung cancer, to increasing literacy rates in rural areas, to conserving endangered species. These examples are just scratching the surface—AI can help driverless cars maneuver road barriers, sanction loans to the needy, detect fraud in real-time, save lives through natural disaster mitigation with flood forecasting in advance and research on predicting earthquake aftershocks, to name a few.

As AI expands our reach into the once unimaginable, it also sparks conversation around topics like fairness and privacy. Fairness, Privacy, Safety & Security are important conversations related to AI implementations that requires the engagement of societies globally. That is where the benefits of adopting XAI techniques become extremely useful, due to their powerful ability to describe black-box models in an understandable, hence conveyable fashion towards stakeholders, who come from various multi-disciplinary background.

XAI can effectively ease the process of explaining the reasons why a model reached a decision in an accessible way to non-expert users, i.e., the rationale explanation. This confluence of multi-disciplinary teams in projects related to Data Science and the search for methodologies to make them appraise the ethical implications of their data-based choices is absolutely critical. It is in this field where XAI can significantly boost the exchange of information among heterogeneous audiences about the knowledge learned by models.

To sum up, the advantages of Explainable AI are the following three major points:

a. It helps the Data Scientists, those who builds AI algorithms in the first place, to debug and optimize the models and design interpretable and inclusive AI systems from the ground up with tools designed to help detect and resolve bias, drift, and other gaps in data and models. It also helps them to grow end-user trust and improve transparency with human-interpretable explanations of machine learning models thus being able to deploy AI with confidence.

b. It helps in streamlining Model Governance, which is becoming a key function within an enterprise, as more and more organizations are resorting to AI/ML-based interventions in their business processes and outcomes and also offering AI/ML powered products and services to their clients that can impact the society and human lives.

c. Ultimately, Explainable AI (XAI) is a step towards Responsible AI.

4 Conclusion

As we saw above, Explainable Artificial Intelligence (XAI), is emerging as a vital force behind the adoption of ML/AI methods in real-life applications. Fundamentally, the concept of explainability, needs to be approached from three different perspectives, as we saw above:

1) Explainability vs. accuracy trade-off

2) Explainability by design i.e. using ML/AI models that feature some degree of transparency, thereby interpretable to an extent by themselves; and

3) Post-hoc XAI techniques devised to make ML/AI models more interpretable especially for individual instance of predictions by the model.

The field of XAI is still evolving. There are still some limitations to overcome. While XAI technique for shallow models based on structured data have reached a reasonable level of maturity, the area of explainability of Deep Learning models and handling unstructured data like images and media files, is still an area under research in both industry & the academia.

Of late, any discussions on XAI seems to tread in the realm of Responsible AI, a paradigm that imposes a series of AI principles to be met when implementing AI models in practice, including fairness, transparency, privacy, accountability, ethics, security and safety. The vision of XAI, as a core concept, is to ensure the aforementioned principles for Responsible AI are uphold to the hilt. While the Data Scientist community will be largely responsible in realizing this vision, academicians, the tech community, the regulators, the business community will also have to be equally accountable.

References:

1. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, By Christoph Molner, Website - https://christophm.github.io/interpretable-ml-book/
2. The Importance of Human Interpretable Machine Learning - A brief introduction into human interpretable machine learning and model interpretation, By Dipanjan Sarkar, Website - https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

About the Author

Arup Duttaroy

Dy. Head- Mosaic, LTI

Arup heads the Sales Engineering function of LTI's Mosaic platform. He comes with 25+ years of prior industrial experience with last 20+ years in the Information Management & Business Analytics area. In his business development role, Arup has been spearheading the Go-To Market Strategy of LTI's Mosaic Product suite and have been instrumental in closing multiple deals for customers across the globe. He is a Gold Medallist, being the topper of the PGDBM (Evening) batch 1993 at XLRI, Jamshedpur, India, specializing in Finance & Systems.