lumin

by Fosfor

Whitepaper

Demystifying Explainable AI for Business Decision Making

Author: Shalini Harkar,
Senior Data Scientist,
Lumin by Fosfor

Table of Contents

Abstract

As Artificial Intelligence becomes increasingly part of our daily lives, the predictions being made by AI-enabled systems are becoming much more profound, and in many cases, critical to life, death, and personal well-being. The need to trust these AI-based systems with all manner of decision-making and predictions is paramount. What subsists at the core of these decision-making systems are algorithmic “black boxes” that stretch the human capability to comprehend the rationale of model decisions. Deciphering these black boxes becomes overly critical for soldering trust and confidence in AI systems and making them reliable and legitimate. Explainable AI (XAI) has the capabilities to overcome these concerns, while providing reassurance that decisions are made in an appropriate and non-biased way.

In this paper, we look at how Explainable AI frameworks and interpretability techniques make AI-driven decisions more transparent, reliable, and trustworthy for business users and understand the key determinants and uses of explainability in business decision-making.

Introduction

Explainable AI – Unveiling the AI 'Black Box'

The promise of what AI can do for organizations has been at a fever pitch over the past few years. Continued improvements in AI systems have produced autonomous machines that learn and act on their own. All these advancements provide enterprises with an enormous opportunity to improve operational efficiency and drive profitable growth across domains such as financial services, insurance, healthcare, and human resources.

The future of AI lies in enabling people to collaborate with machines to solve complex problems. Like any efficient collaboration, this requires good communication, trust, clarity, and understanding. However, AI and Machine Learning models and their outcomes can be non-intuitive and difficult for people to understand, which leaves many feeling unsure about what their AI is actually doing. A rapid surge in the complexity and sophistication of such AI-powered "black box" systems has evolved to such an extent that humans do not understand the mechanisms by which AI systems work or how they make certain decisions, especially when it comes to computing outputs that are unexpected or seemingly unpredictable.

Explainable Artificial Intelligence (XAI) refers to methods and techniques that produce accurate, explainable models of why and how an AI algorithm arrives at a specific decision so that AI solution results can be understood by humans (Barredo Arrieta, et al., 2020). XAI provides the needed understandability and transparency to enable greater trust toward AI-based solutions. The model explanations are typically extra metadata information in the form of some visual or textual guides that offer insight into specific AI decisions or reveal the internal functionality of the model as a whole.

With explainability, the AI technology assists human beings in making quick, fact-based decisions but allows humans the capability to still use their judgment to assess or add the context in which the decision is to be taken. It's critical for organizations to understand how their models make decisions because

- It makes it easier to explain the findings or outputs to non-technical audiences so they can make confident decisions
- It provides the glimpse of the intended impact or pre-empts any negative or unforeseen consequences for data practitioners and thus gives them a chance to course-correct and fine-tune their models in production

Whether by pre-emptive design or retrospective analysis, new techniques are being employed to make the black box of AI less opaque. Using Explainable AI, businesses gain greater visibility over unknown vulnerabilities and flaws as well as the reassurance for stakeholders that the system is operating as desired.

Gauging the Need for Explainability in AI

As the move to Explainable AI gains pace and momentum, it also opens up the conversation around the need and extent of explainability required for taking effective business decisions.

Explainable AI looks at why a decision was made by an AI system, so that the outputs can be more interpretable for human users. It enables them to understand why the system arrived at a specific decision or performed a specific action. XAI helps bring transparency to AI, potentially making it possible to open up the "black box" and reveal the full decision-making process in a way that is easily comprehensible to humans.

This leads to an interesting question- Is there a need for explainability for every AI-based decision for it to be trusted?

While several factors like greater need for transparency, trust and higher adoption of AI-based applications are propelling the need for better explanation, not all AI techniques require full explainability. A fundamental litmus test for explainable AI is the greater the potential consequences of AI-based outcomes, the greater the need for Explainable AI.

The underlying dynamic is that, as the model complexity and potential impact increases, so does the need to explain.

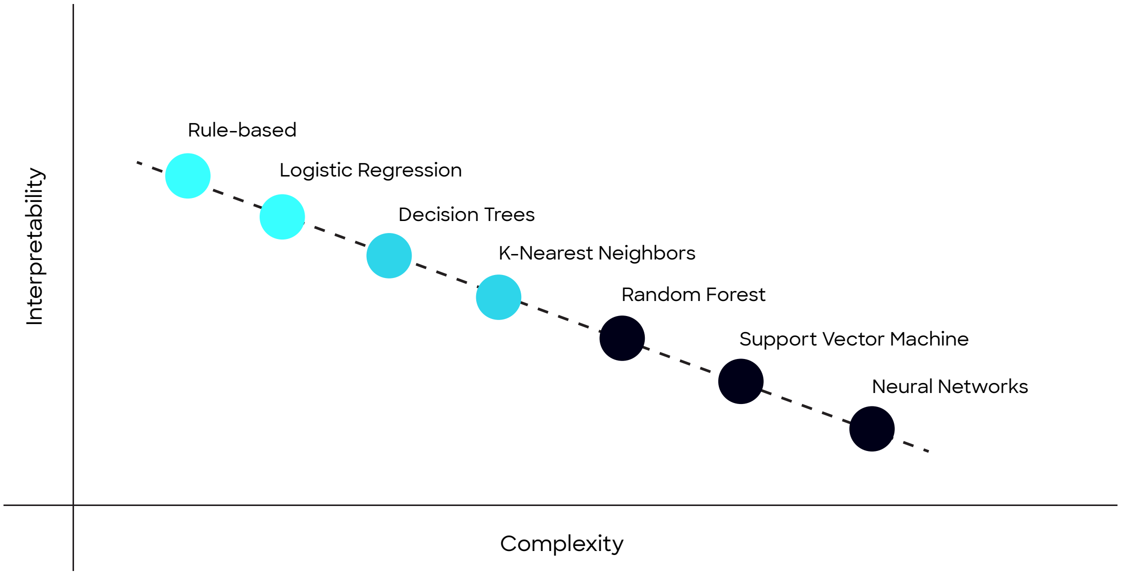
Pre-requisites for AI explainability

- **Degree of Algorithmic Complexity –** An important outlook for comprehending the need for Explainable AI emerges from the model's complexity perspective. "Black box" AI models have unobservable input-output relationships and lack clarity around their inner workings. This is typical of deep learning and boosted / random forest models which model incredibly complex situations with high non-linearity and interactions between inputs. There are no clear steps outlining why the model has made the decisions that it has, so it's difficult to discern how it reached the outputs and predictions that it did. Generally, simpler models like linear or rule-based models are straightforward to interpret but as the model complexity shoots up, it gets equally hard to comprehend. Hence, the need for model explainability becomes inevitable.
- **Criticality of Use Case –** Necessity for AI-explainability becomes indispensable for the applications where the AI decision has a striking impact on humans and society at large. For applications like AI-powered chatbots or recommendation engines for e-commerce portals, it doesn't really matter if the AI system operates in a black box. But for use cases with a big human impact – autonomous vehicles, aerial navigation, drones, or military applications – being able to understand the decision-making process is mission-critical. Medical diagnosis, loan approval, recruitment processes, legal justice, and many more instances necessitate algorithmic transparency and trust.

Uncovering Explainable AI frameworks - Going Beyond Human Explanations

When explaining a concept, humans rely on their rich and expressive vocabulary as well as past experience. But when it comes to critical decisions that either require a lot of data processing or involve multiple interrelated factors, human explanations are often flawed and prone to oversimplification and cognitive biases. In the field of AI, elementary ML models like linear regression and rule-based models are straightforward for humans to interpret due to their intuitive nature, so explaining them becomes less challenging. With an increase in model complexity, it becomes equally hard to bring explainability without having robust Explainable AI frameworks in place.

Why Explainable AI



When interacting with algorithmic decisions, users now expect and demand expressiveness from AI systems. An applicant whose loan was denied by a bank's AI system will want to understand the main reasons for the rejection and what she can do to reverse the decision. A developer may want to understand where the model is lacking as a means of improving its performance.

Let's take a closer look at how different Explainable AI frameworks provide a more accurate way of comprehending models compared to human interpretation for various use cases

- **Explaining Multiple Feature Interactions -** Financial institutions like banks use AI systems to make critical decisions like approving or denying loans to customers. Let's assume the AI system has used a tree-based model to predict the loan approval or denial for an applicant. Comprehending tree-based rules that decide the model outcomes do not contemplate feature interactions between the variables. Feature interactions are a prime aspect of decision-making and cannot be disregarded.

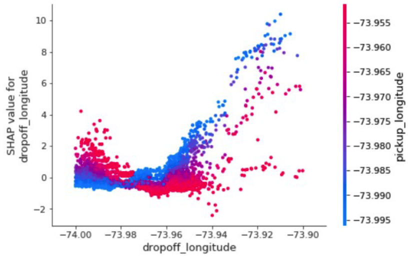
  Shapley value, a solution concept based on coalition game theory provides an excellent approach to explain model's outcomes considering the contribution of each feature in tandem. For example, in sales forecasting, Shapley value decodes whether the contribution of an attribute is significant and should be given priority or not. Features with high Shapley value point out that its contribution for driving sales is significant and should be outweighed while features with a low Shapley score clearly hint that its contribution powering sales is not significant and should be neglected.

- **Assessing Impact of Features on Model Outcomes -** Deciphering how a change in one feature impacts the outcome is difficult to comprehend by humans. A framework like **Partial Dependence Plots (PDP)** provides a visual interpretation to show marginal changes in a model's output (predicted response) when a feature is changed to understand the dependence of features on the target outcome.

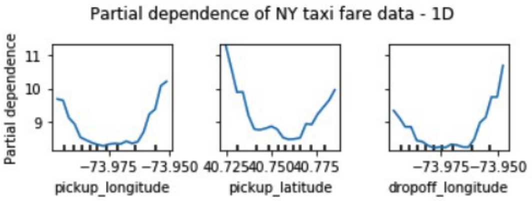
- **Testing Causal Relationships on Model Outcomes -** Explainable AI frameworks also answer many other questions that humans cannot comprehend just by inspecting the model. For example, in case of a denied loan application, it is important for the bank to let the customer understand the reasons for denial and what can be done to reverse the decision.

  Counterfactuals, an Explainable AI approach provides a way for the applicant to understand what feature or aspect can be improved to get the application accepted. So, a statement like "You were denied a loan because your annual income was $30,000' when shared along with an actionable recourse like 'If your income had been $45,000, you would have been offered a loan" gives a clear explanation of what's required. Here, the second sentence is a counterfactual explanation.

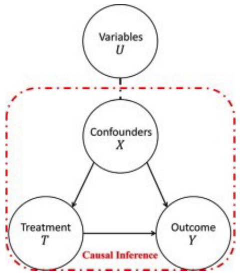
- **Understanding Data Bias and Model Fairness -** It is difficult for humans to discern whether the AI-based decision for loan approval or denial is biased in any way basis the race, gender, caste, or occupation of the applicant. Various algorithm fairness metrics at hand can help to trace if the decisions are fair or not.



[1]Example of Shapley Plot



[1]Example of Partial Dependance (PDP) Plot



[2] Graphical Representation of Counterfactual explanation

Source- [1]Towards data science, [2] Sciencedirect.com

There are many other factors that could contribute to how an AI model operates and makes its predictions, and thus, many ways to explain them. Apart from model explanation, model accuracy is also a key factor in assessing its validity and robustness. Prepending model accuracy with fair explainability pulls out the algorithmic opaqueness and adds glassiness to the decisions.

Pre-modelling and post-modelling explainability are two stages which can bring confidence in AI-driven decisions.

Pre-modelling explainability is a collection of diverse methods with a common goal of gaining a better understanding of the dataset used for model development. This approach is motivated by the fact that the behaviour of an AI model is, to a large extent, driven by the dataset used to train it. These could be solved by

1. Identifying Hidden Bias in the Data

Bias in the data can be induced unintentionally by humans due to which the sample is not representative of the entire population and hence a biased dataset results in inaccurate accuracy levels, skewed and unfair outcomes. Identifying the type of bias and its mitigation preparatory to model development becomes paramount. This step ensures that quality of data fed to the model is equitable and non-biased for model training.

2. Exploratory Data Analysis

The goal of exploratory data analysis is to extract a summary of the main characteristics of a dataset which includes various statistical properties of the dataset such as its dimensionality, mean, standard deviation, range, missing samples, and so on. Real-world datasets are often complex and high-dimensional. Visualizing such high-dimensional data can be a challenge, as humans can only imagine up to three dimensions with ease. This could be solved using parallel coordinate plots, or dimensionality reduction methods like principal component analysis (PCA) and t-SNE to gain insights about the data prior to model building.

3. Prescriptive Data Analysis

Prescriptive analysis helps deduce inferences from the respective data using population metrics, sample statistics, confidence interval to deduce conjecture about a population.

Post-modelling Explainability
Post-modelling explainability describes the rationale behind the model selection, unmasks the model parameters at which model has performed, or unveils model reliability.

Achieving cognizance at the data as well as the model level using these techniques brings unabridged comprehension to the end user and helps them make trustworthy decisions.

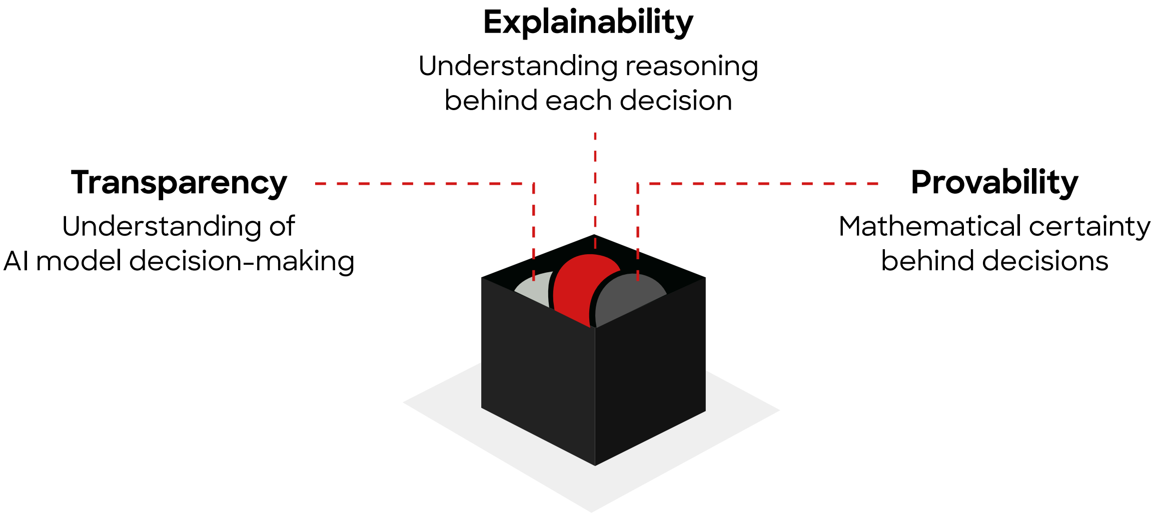
Explainability in Business

Providing explanations to the business process, its decisions, and activities is an important factor to minimize and deal with the ambiguity of the business process that causes multiple interpretations, as well as to engender the appropriate trust of the users in the process. While the equations that make up Machine Learning algorithms are often straightforward, it can often be difficult to derive a human-understandable interpretation for how latent features in a large dataset have been weighted. A model is better interpretable than another model if its decisions are easier for a human to comprehend. Higher the interpretability of a Machine Learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made.

In a technical context, AI interpretability is applied to the process of examining rules-based algorithms, while explainability is applied to the process of examining “black box” deep learning algorithms. In AI UX contexts, the distinction between explainability and interpretability relates to how AI problems are presented to different types of users or as part of different workflows.

- For users, providing valid explanations for an AI or Machine Learning model boosts business user confidence in the reliability of the models
- For developers, providing a high degree of transparency into AI and Machine Learning models is critical in helping developers defend the validity of their models and in providing explanations to decision-makers
- For C-suite members, transparency in AI / Machine Learning models helps accountability and adherence to regulatory practices

Three Keys to Understanding AI Decision Making



Explainability
Understanding reasoning behind each decision

Transparency
Understanding of AI model decision-making

Provability
Mathematical certainty behind decisions

Source: PwC

For any explanation to have value and be effective in business context, it should fulfil the following five measures

1. **Context**
   Who is receiving the explanation, what is the intended use case and potential impact, what is their level of expertise, and what are their time constraints

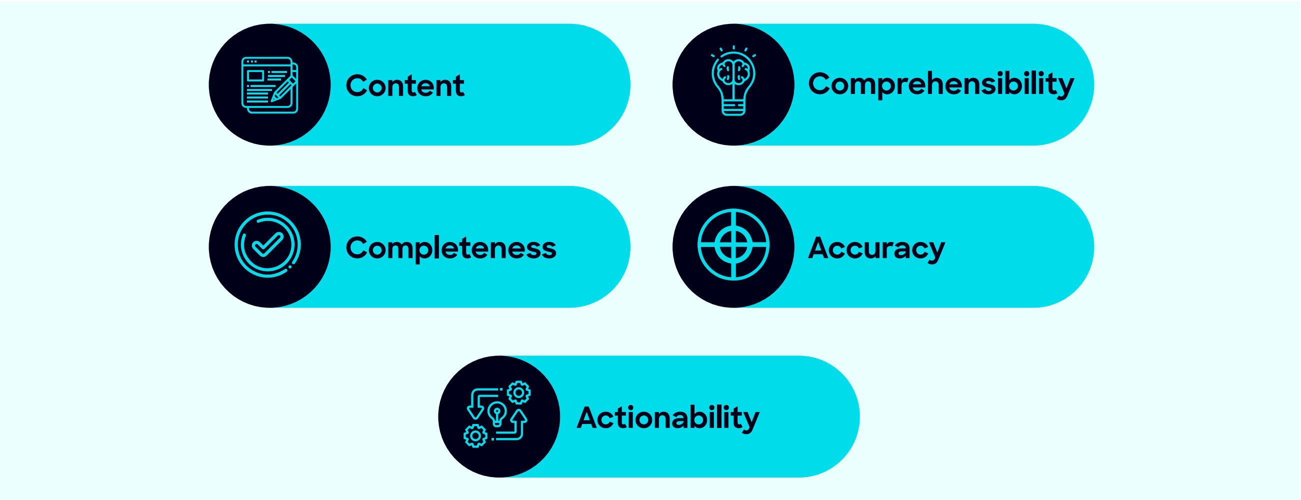
2. **Comprehensibility**
   How much effort is required for technical or non-technical users to interpret it

3. **Completeness**
   if the 'explanation' explain the decision completely or only partially

4. **Accuracy**
   How accurate is the explanation and what other models were used for the decision

5. **Actionability**
   How actionable is the explanation? What can we do with it?

Explainability in Business



Content

Comprehensibility

Completeness

Accuracy

Actionability

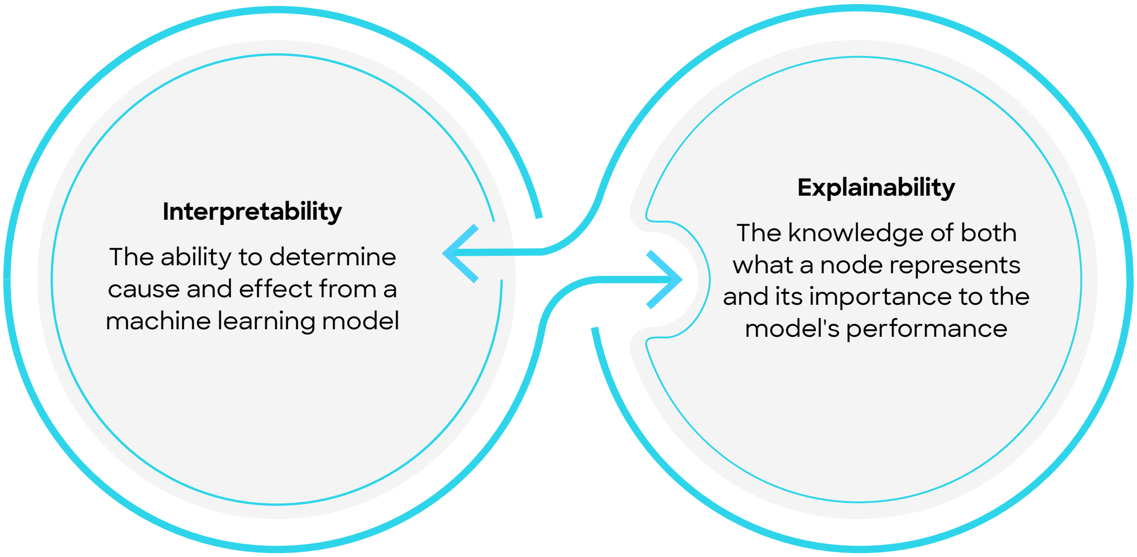
At the end, explainability is directly proportional to the availability of data and context, together with the algorithms that are employed and modified. For each AI use case, the verification and validation process may require different elements of interpretability, depending on the level of rigor required. The business assessment and cost analysis is also required to determine the level of explainability that should be provided to users over and above the use case implications.

The Interpretability Versus Accuracy Trade-off

When attempting to explain the predictions of Machine Learning (ML) models in the real world, it is critical to be precise. One of the main motivations behind explainability techniques is to build trust in ML models. If the explanations are wrong, this trust is broken.

Explainability comprehends the internal working mechanics of an underlying model and what makes an effective explanation for the human users of an AI system. Interpretability, on the other hand, gauges how well any user, especially a non-expert, would understand why and what the system is doing either overall or for a particular case. Based on this notion, the term explainability and interpretability is often used interchangeably, despite immense differences in intention and practical application.

Interpretability enables transparent AI models to be readily understood by users of all experience levels. Explainable AI applied to “black box” models enables data scientists and technical developers to derive an explanation as to why models behave the way they do. Most business stakeholders demand inherent interpretability to take critical decisions but this in turn restricts modelling and can lead to lower accuracy. A naturally arising question is whether there are some inherent trade-offs between the “interpretability” of an algorithm and its potential power in terms of accuracy or performance.


Interpretability
The ability to determine cause and effect from a machine learning model
Explainability
The knowledge of both what a node represents and its importance to the model's performance

For a given problem, it is critical to have a clear idea of what is the priority - accuracy or explainability so that this trade-off can be made explicitly rather than implicitly. In medical diagnosis, detection of malignancy best exemplifies this notion. In practice, the patient being diagnosed with a suspected case of cancer is told that the most precise AI model has detected that the tumor is non-malignant with model accuracy >90%. A vexed question to answer is how fair and trustworthy is the model's prediction with absolutely no clue about the techniques or process by which the model arrived at the output?

Explainable AI frameworks spell out the why and how of the model that determined the non-malignancy by elucidating the features and their contribution to the prediction in a user-friendly manner.

Ultimately, the discussion about interpretability versus explainability should start with why interpretability and explainability are important for various individuals followed by the robustness of the model. To put interpretability into use, evaluation is a must. Evaluation at the human level, function level, and application level are to be considered before relying on interpretations.

There are some cases in which an interpretable model can be made more accurate by lengthy feature engineering. These cases require a commercial decision on whether the need for complete explanations is high enough to outweigh the costs associated with the feature engineering process.

Explainability in Action – Lumin's XAI

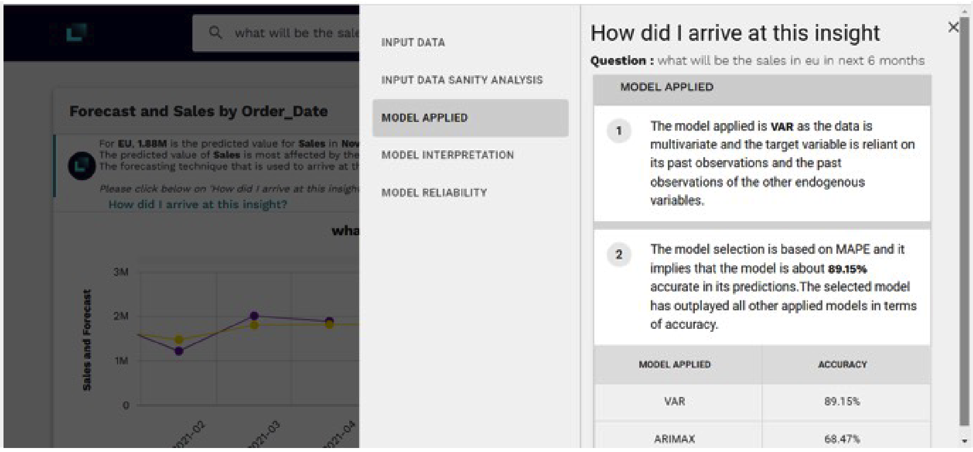
There are significant business benefits of building interpretability and explainability into AI systems. As well as helping address pressures such as regulation, and adopt good practices around accountability and ethics, there are significant benefits to be gained from being on the front foot and investing in explainability today.

Explainability is at the core of Fosfor Lumin's breakthrough AI, NLG, and cognitive analytical engine. Here are some of the use cases of how businesses are using Lumin's XAI to enhance their AI systems, ensuring accuracy at one end and intended impact on the other.

- Lumin's explainable AI capability offers an efficient way to understand how the historic trends impacts the future projections. Working as a clinical decision support system for a leading Fortune 500 pharma major, Lumin's XAI has turned down the cost of error and brings reliability to the end-user for more adequate utilization. For example, the cost of error in the event of misclassification of malignancy is now overcome by apprising the rationale behind the misclassification
- In Retail, accurate sales forecasts empower organizations to make smarter business decisions, and uncovering how sales are influenced by the historical trend helps to achieve a clear competitive advantage. Lumin's explainable AI capability offers an efficient way other than standard methods for the end-user to decide which attributes drive the sales so that business actions could be taken in accord. For global CPG major, Lumin's XAI is helping sales leaders with interpretable metrics to focus on relevant attributes driving the sales

  The intuitive analysis and step-by-step explanation of the modelling techniques has enhanced the overall sales performance score and has helped the team understand the nuances and implications of factors that impact sales

Lumin's Explainable AI – How Did I Arrive at This Insight?



For **EU**, **1.88M** is the predicted value for **Sales** in **Nov**
The predicted value of **Sales** is most affected by the
The forecasting technique that is used to arrive at th

Please click below on 'How did I arrive at this insigh

How did I arrive at this insight?

How did I arrive at this insight

Question : what will be the sales in eu in next 6 months

MODEL APPLIED

1 The model applied is **VAR** as the data is multivariate and the target variable is reliant on its past observations and the past observations of the other endogenous variables.

2 The model selection is based on MAPE and it implies that the model is about **89.15%** accurate in its predictions.The selected model has outplayed all other applied models in terms of accuracy.

MODEL APPLIED	ACCURACY
VAR	89.15%
ARIMAX	68.47%

Conclusion

Gartner reports that by 2025 “black box-driven” decision-making will have a colossal societal concern. Going forward, as AI promises to become more autonomous and help identify dangerous industrial sites, warn of impending machine failures, recommend medical treatments, or take countless other critical decisions, the need for high-quality explanations will be paramount. To make this possible, Explainable AI needs to meet the demands for understandable, transparent, interpretable, and (consequently) trustworthy AI-based solutions. It’s the appropriate time now to discern the urgency to develop AI-solutions ingrained with explainability.

XAI plays a fundamental role in gaining human operator trust on one hand while supporting established guidelines, standards, and regulations on the other. Finding the right equilibrium between accuracy and interpretability of underlying algorithms are increasingly important in business context to maintain trust and transparency in AI systems without severe cost implications. Hence explainable AI is a pre-eminent route for responsible AI serving in resolving ambiguity where decisions drawn from systems are unforeseen.

References

- A Unified Approach to Interpreting Model Predictions by Scott M. Lundberg eta.al, 31st Conference on Neural Information Processing Systems (NIPS 2017).
- Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) by Amina Adadi eta.al, IEEE Access, 2018.
- AI Fairness by Trisha Mahoney, Kush R. Varshney, Michael Hind.
- Explainable AI Driving business value through greater understanding, PwC https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf
- S. Wachter, B. Mittelstadt, and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology, Volume 31, Number 2 Spring 2018, p. 844.
- https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf
- https://searchenterpriseai.techtarget.com/feature/UX-defines-chasm-between-explainable-vs-interpretable-AI
- https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends

By Shalini Harkar

Senior Data Scientist
Lumin by Fosfor

Shalini holds a Masters in Human Genetics from the University of Jammu. She has 8+ years of experience across academia, industry in Healthcare Analytics, delivering high business impact solutions in healthcare. Her core expertise lies in data and AI research and development with multiple research handouts published in various scientific international journals.

In her current role as Senior Data Scientist at Lumin, Shalini has been working towards forging a new way of business decision -making by bringing explainability in AI-enabled decisions.

The Fosfor Product Suite is the only end-to-end suite for optimizing all aspects of the data-to-decisions lifecycle. Fosfor helps you make better decisions, ensuring you have the right data in more hands in the fastest time possible. The Fosfor Product Suite is made up of Spectra, a comprehensive DataOps platform; Refract, a data science and MLOps platform; and Lumin, a decision intelligence platform. Taken together, the Fosfor suite helps businesses discover the hidden value in their data. The Fosfor Data Products Unit is part of LTIMindtree, a global technology consulting and digital solutions company with hundreds of clients and operations in more than 30 countries. For more information, visit Fosfor.com.

FOSFOR