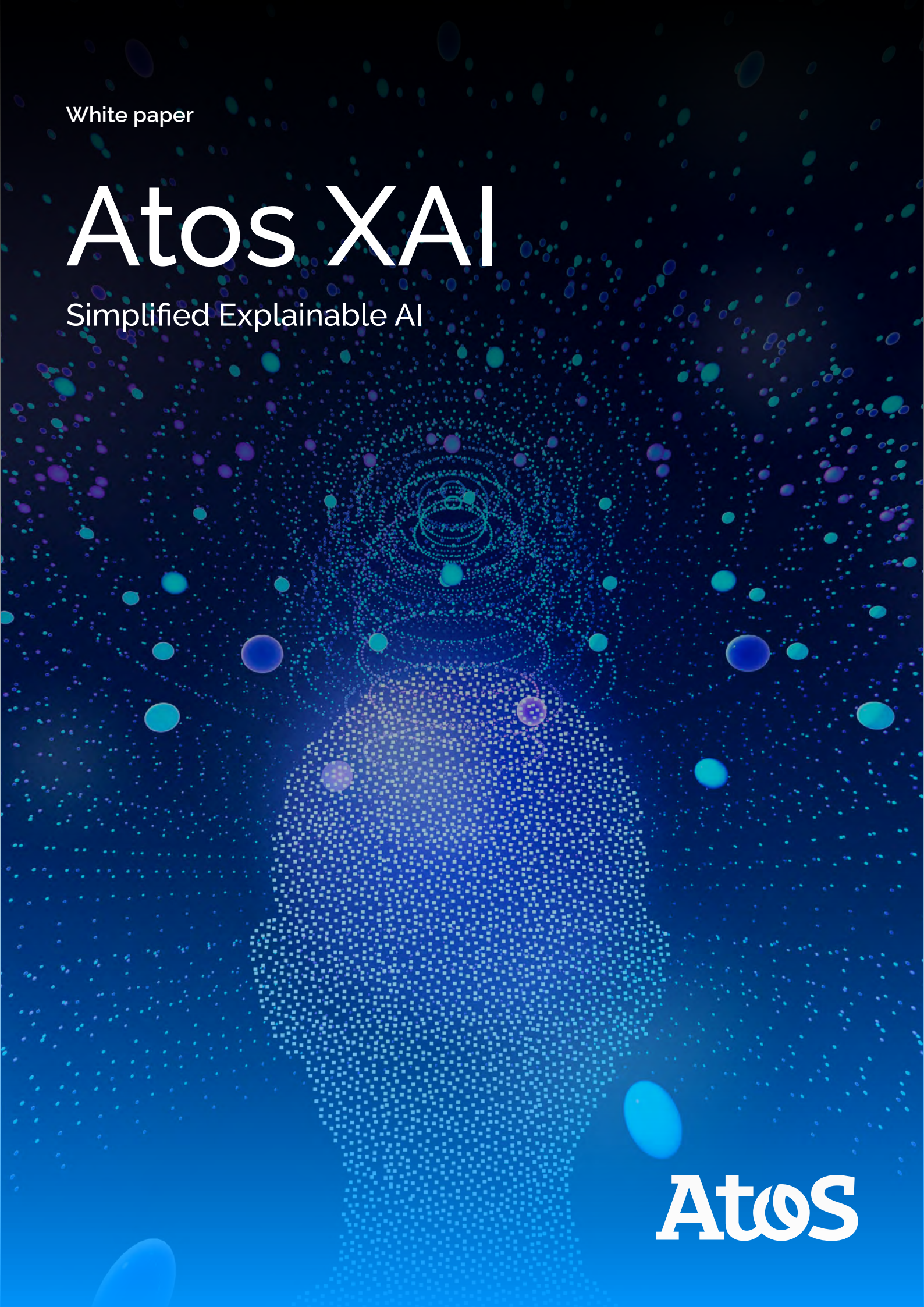
White paper

Atos XAI

Simplified Explainable AI



Atos

As artificial intelligence (AI) becomes entrenched in mainstream applications, regulations and laws are being written that lay emphasis on managing the associated ethical and operational risks. The need to understand why an AI-based system made a particular recommendation is critical to ensuring recommendations are fair and beneficial to the business.

In response, Explainable AI (or XAI) has emerged as a digital discipline verifying that an AI-based system delivers expected results that are interpretable. XAI is attracting significant R&D investments, with the intent of making ML-based solutions explainable and trusted.

Several explainability solutions are available for ML models, however, most suffer from the drawbacks including:

- Solutions/methods are overly technical and require some degree of specialization to interpret results
- Solutions do not deliver a comprehensive or varied explainability view.

Atos XAI addresses these challenges by:

- Providing comprehensive but easy-to-understand reporting, focusing on the explainability of the ML model in terms that users without a technical background can understand
- Hiding underlying complexity so users are presented with a simple interface to create explainability reports for their ML models.
- Providing unique perspectives of model explainability using different algorithms for the same ML model.

Integrating Atos XAI into most solutions expands the opportunity to introduce Explainable AI to more digital ecosystems in the organization.

Contents

Introduction

Recent advances in AI-based systems have led to organizational dependence on AI/ML to optimize operations or help stay ahead of the competition. AI-driven decision-making is now ubiquitous and an integral part of our day-to-day life, regardless of industry, and helps make critical decisions in many sectors. This may lead to legal and regulatory compliance implications. Think of self-driving car decisioning, choosing candidates to call for a job interview or predicting fraudulent banking transactions — and the risks and consequences should there be no regulation. It presents us with the conundrum: If a machine learning (ML) model makes correct decisions, should we not just be happy with the result and ignore its rationale for making those decisions?

In the real-world scenarios encountered by ML models, predictions and model accuracy only solve part of the problem. The system must also account for why the prediction was made, as the decision may have severe implications for those affected by it. Explainability in ML systems presents the rationale for the decisions in a more understandable form.

Simpler algorithms, like linear and logistic regression, are best suited for explainability, and underlying formulas and mechanisms are easily understood. Still, in many real-world use cases, these algorithms cannot supply adequate performance metrics. Hence it becomes exceedingly difficult to use these models in production.

"The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks."

(Doshi-Velez and Kim, 2017)

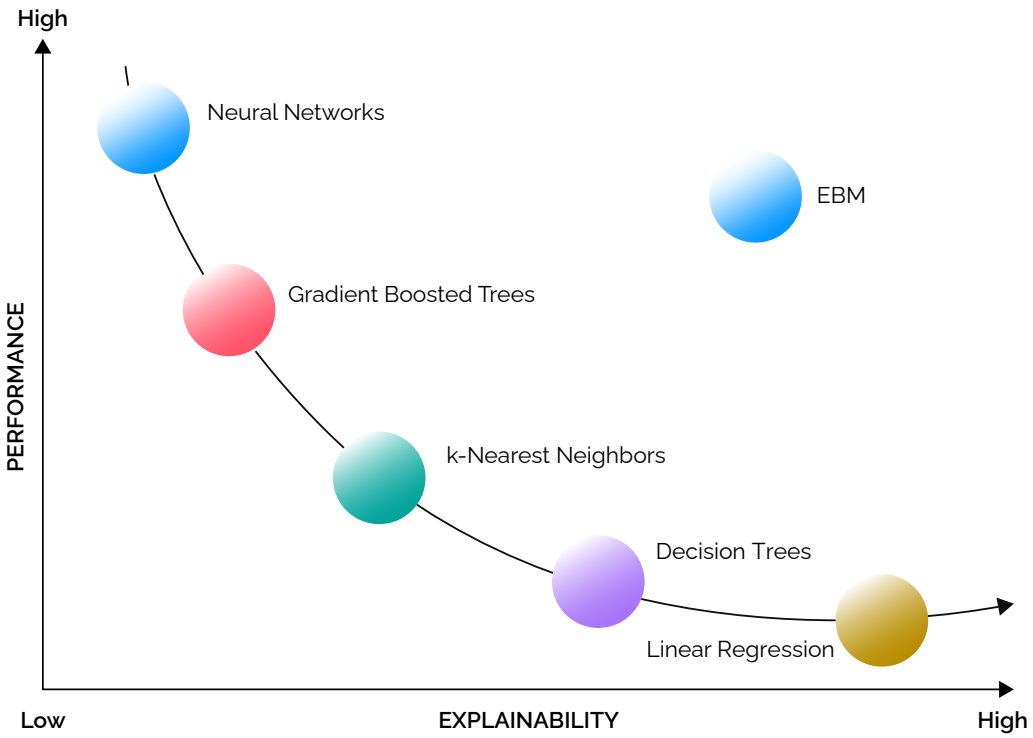


Fig 1.0 - **Performance and explainability with EBM (Picard 2021)**

To improve accuracy, precision and recall, we use complex ML and deep learning (DL) algorithms like Extreme Gradient Boosting (XGBoost), support vector machines (SVM), stacking and neural networks. These advanced algorithms can generate better results, but at the cost of losing explainability. These algorithms and their mathematical calculations are so complex that they cannot be understood or explained easily on their own. Fig 1.0 shows the relationship between explainability and model performance.

Here we can see that the complex algorithms give better performance, but they're black boxes; we cannot see what steps the algorithms follow to make a prediction. The results may be fascinating, but due to the black-box nature of the algorithms, it is almost impossible to trust these algorithms. Once we understand the reasoning behind a particular algorithm, we can trust the predictions and deploy to production with confidence. With the predictions now easily interpreted, they can be used to make better decisions.

Explainability can also help us find biases in the data and the models. For example, in a loan-approval ML use case, explainability could reveal that our model may be declining loan applications for a specific group of individuals based on their demographic or ethnic origin.

Some of the benefits provided by model explainability are:

- Building trust in ML/DL models
- Increasing transparency
- Identifying bias in data or ML/DL models
- Improving troubleshooting resolution.

Many times in technology's history, we've seen that early and broad adoption can lead to differentiating value, growth and competitive advantage. Atos has found a way to democratize the creation, accountability and distribution of AI models for impactful effect.

XAI Framework Design

Let's review the critical methods in XAI.

ML explainability can be intrinsic or post hoc. Intrinsic methods restrict the complexity of the models and are considered interpretable due to their simple structure, such as short decision trees or sparse linear models. Post hoc interpretability refers to applying interpretation methods on top of trained models.

Similarly, *local explanation* explains the model at record level, and *global explanation* explains the model at feature level.

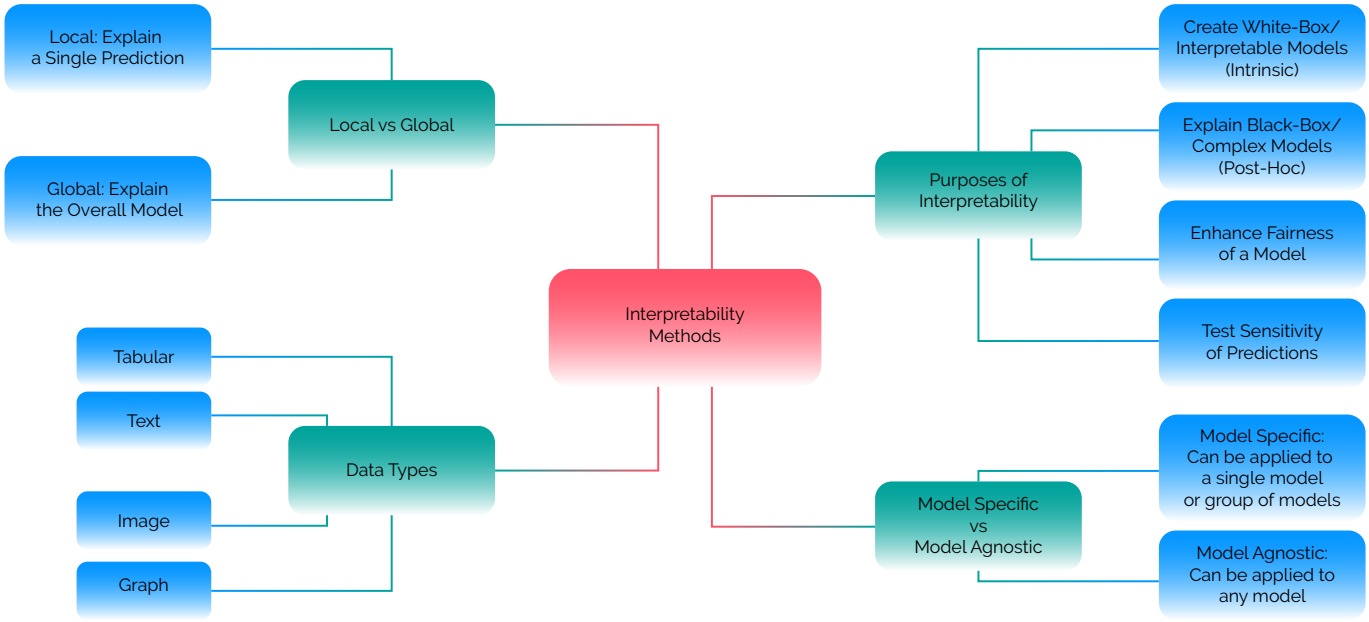


Fig 2.0 - **Mind map of XAI/Interpretability methods. (Linardatos et al., 2020)**

Atos XAI Framework

There are different methods available to derive ML explainability. Each comes with its own complexity and learning curve.

At Atos, we developed a comprehensive framework that reduces the underlying complexity while exposing application programming interfaces (APIs) to expand utilization. This simplified API helps ML engineers and data scientists perform experiments on their models to derive interpretability at both the model level and the record level. This framework is extremely easy to configure and helps to derive a quick and qualified explanation of the model and prediction, producing a simplified, easily understandable report. The XAI report gives a unified view of the model interpretation using different methods or algorithms. It improves the trust and reliability of the model.

The XAI report also creates a simple natural language-based explanation of each of the graphs and diagrams. Additionally, the rule-based explanation uses a Fold algorithm to help interpret the rules and exceptions for the model based on the data used to train it.

Our automated ML for explainability solution, XAI, provides a comprehensive report with various explanations at both global and local levels.

Below are the salient features of Atos XAI:

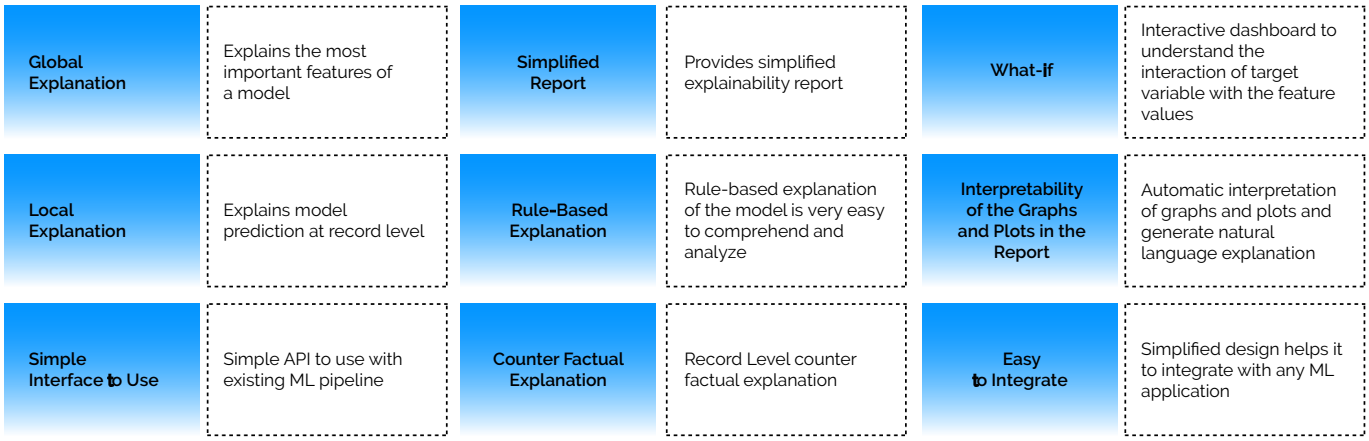


Fig 3.0 - Features of Atos XAI

Report Layout of XAI

Atos XAI is a Python-based solution that provides a comprehensive and easy-to-understand report.

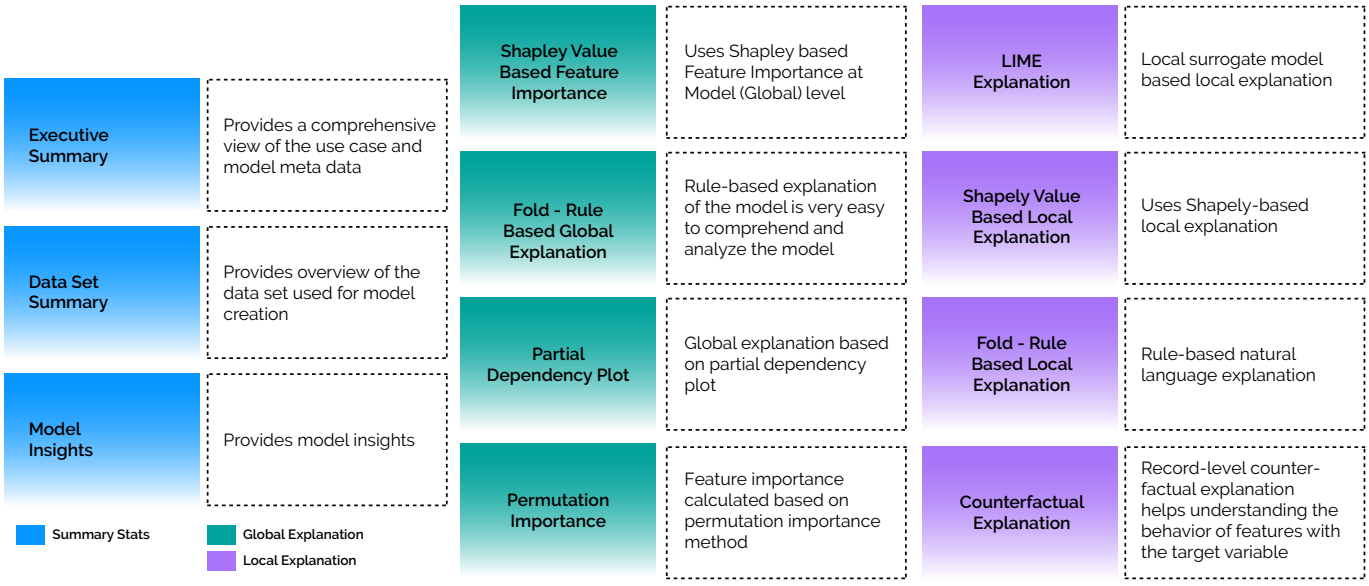


Fig 4.0 - Features of Atos XAI Reports

Use Case Executive Summary

An overview of our use-case experiment, offering a snapshot of dependent and independent features and the essential elements of the use case.

Model Insights

Insights into the topmost relevant features from the experiment that drive model prediction. Different algorithms are used to identify the topmost prominent features.

Dataset Summary

An overall summary of our dataset and features. Details including several observations, target features, model name, operation type and others are included.

Model-Level Explanation

Global explanation of the model is derived over the entire dataset. It consists of the following sections:

1. Permutation Feature Importance

Provides us with information about the importance of each feature on the model prediction by calculating the increase in the model's prediction error after shuffling the value of a feature.

A feature is deemed important if shuffling its value increases the model prediction error. If the error is still the same or decreases, we assume the feature is not essential to the model.

2. SHAP Feature Importance

The Shapley value derives its concept from game theory to measure how much an individual player contributes to a game. To achieve this, we observe each group of players and the outcome they achieve. A player's Shapley value is the weighted average difference between the coalitions that include the player and those that don't include the player. In the case of a model prediction, each player becomes a feature and the game becomes the prediction of the model.

The SHAP, abbreviated from SHapley Additive exPlanations depicts the feature importance along with its correlation, i.e., whether prediction probability increases or decreases with an inclusion of a feature.

3. FOLD Explanation

First-order learner of defaults (FOLD) algorithm uses high-utility itemset mining and answer-set programming logic:

- The FOLD algorithm learns a concept as a default theory (also known as rules) and a set of exceptions.
- These rules and explanations are logical explanations that are easy to understand.

The FOLD algorithm provides explanation in terms of rules and sets of exceptions, which are remarkably like how human being perceives and understands real-world experiences .

Problem Statement
Is a loan applicant eligible for a loan?

The FOLD algorithm uses the data and model to create natural language-based rules and exceptions, as explained below:
Loan applicant is suitable for a loan if the rule and exception combination is true.

A sample of rule and exception explanation is as below:

(Here we are showing two rules, but there can be greater numbers of rules and exceptions)

Rule 1:

Credit score is greater than 750 and

Exception 1 is false

Rule 2:

Credit score is greater than 650 and

Annual income is greater than 100,000 and

Exception 2 is true

Exception 1:

Current loan amount is greater than 700,000

Exception 2:

Months since last delinquent is greater than 60 and

Number of credit problems is less than 2 and

Years of credit history is greater than 10

4. Partial Dependence Plot

The partial dependence plot (PDP) shows the marginal effect one or two features have on an ML model (J. H. Friedman 2001). PDPBox can visually decide the impact of a particular feature on model prediction. It can even help explain the non-monotonic effect of features on prediction.

PDP explains each feature and its impact on output probability with a change in the feature value.

Data-Level Explanation

This section can also be referred to as Local Explanation because it supplies the explanation about each data point or record. It consists of the following sections:

1. **LIME Local Explanation**
   In local interpretable model-agnostic explanations (LIME) papers, authors propose a concrete implementation of local surrogate models. Surrogate models try to approximate the black-box model we are trying to explain.

   LIME is used to train a local surrogate model to explain individual predictions. It creates a new dataset consisting of perturbed samples and the corresponding predictions of the black-box model. On this newly created dataset, LIME then trains an interpretable model, weighted by the proximity of the sampled instances to the instance of interest. The learned model is a good approximation of the ML model predictions locally, but it does not have to be an excellent global approximation. This kind of accuracy is also called local fidelity.

   The LIME method provides an explanation for each data point or record; it gives a detailed explanation of all the features and their value that the model took into consideration to arrive at the output prediction. It provides the probability value for all the target classes, the feature value and its impact on the prediction.

2. **SHAP Local Explanation**
   A Shapley value-based explanation provides a concise plot depicting the impact of each feature and its value on the prediction.

3. **FOLD Local Explanation**
   FOLD local explanations provide the natural, human-readable explanation, in terms of rules and exceptions, that is extremely easy to comprehend and understand.

4. **Counterfactual Explanation**
   A counterfactual explanation shows us what feature value changes can move the output prediction class from one level/class to another. It can be beneficial when we need to provide recommendation. For example, in the loan approval use case, if a customer's loan application is rejected, this can be used to give feedback on what changes are needed to make the profile eligible for a loan.

5. **What-if Analysis**
   A what-if tool acts as a playground area where anyone can play around with the record values and see their changes to the output prediction. It can be beneficial for troubleshooting and model performance analysis.

Highlights

Some of the unique benefits of Atos XAI framework:

- Automatic interpretation from graphs/plots

  Many times, graphs can be difficult to interpret. We have added dynamic explanations for each report chart so that end users can easily understand them and make better decisions.

- Rule-based natural language explanation

  Explanations are provided in terms of rules and sets of exception in natural language, so users are better equipped to understand and interpret model behavior. These explanations are beneficial for business users with domain knowledge to help understand the impact of each feature.

- Unified view of explanation using multiple algorithms

  Multiple algorithms come together to derive the explainability, which increases our trust in the trained model and its predictions if they all provide the same kind of explanation. It will help us to troubleshoot our model and find bias, if any.

- Custom ranking based on different explanations

  We have developed a custom ranking mechanism, which takes input from various explanation algorithms and provides a final customized ranking of features.

- What-if analysis

  This tool provides an interactive playground for predictions. Users can change the value of features for any data point and then see its impact on the prediction. This helps to refine models and improve troubleshooting.

Conclusion

XAI is quickly gaining importance, with explainability becoming an essential and default feature of AI-enabled solutions. As AI-enabled system adoption increases, the need for a matured AI system to include vital attributes like explainability, security, privacy, robustness, ethics and inclusiveness is critical. These models need to be immune to bias in the data and demand a robust model governance framework

Therefore, AI engineers must consider these principles in the early phases of AI-enabled system development. With these principles in place, we can ensure an explainable AI framework that earns the confidence of the business and community at large.

XAI is a highly critical component of an overarching responsible AI framework. All AI-enabled systems should explain the action/predictions they have made, as they have far-reaching ramifications for businesses and society.

Atos has been committed to the field of XAI, helping organizations understand the importance of explainability and how to integrate it into ML practices: XAI decreases the risk of AI systems while building trust and reliability. Reports generated from this framework can be convenient for data scientists, ML engineers and business SMEs to understand the model's functioning and predictions, and take further actions as needed.

Key takeaways:

- Explainable AI is one of the core components in building a responsible AI system that increases the trust in AI-enabled systems.
- XAI is a responsible AI-based design approach that can increase the trust of AI-enabled systems.
- Simplified reporting of model explanation increases the understanding of AI systems for all the stakeholders of AI ecosystems.

Where to begin:

As a first step, our recommendation is to identify the AI-enabled systems in your organization. This information will allow you to work toward the below action items:

- Advocate for explainability to be at the forefront of any AI decision-making solutions.
- Get internal and external stakeholder alignment around identified risk areas.
- Gather input from your community of interest to properly reflect rule-set outcomes to avoid risk, bias or unintentional consequences.
- Lay out a framework to enable and integrate XAI for your AI systems.

Citations and References

Citations

1. Dolshi-Velez, F., & KIM, B. (2017, February 28). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv. Retrieved June 6, 2022, from https://arxiv.org/pdf/1702.08608.pdf
2. Picard, R. (2021, November 6). *Performance And Explainability With EBM*. OakBits. Retrieved June 6, 2022, from https://blog.oakbits.com/
3. Linardatos, P., Kotsiantis, S., & Papastefanopolous, V. (2020, December 20). *Explainable AI: A Review of Machine Learning Interpretability Methods*. National Library of Medicine PubMed Central. Retrieved June 6, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7824368/

References

1. The University of Texas at Dallas (Prof. Gopal Gupta, Prof. Doug DeGroot, Farhad Shakerin, Huaduo Wang, Parth Padalkar)
2. *A Guide for Making Black Box Models Explainable*, Christoph Molnar, https://christophm.github.io/interpretable-ml-book/
3. ELI5 Python package repository, https://github.com/eli5-org/eli5/
4. *Greedy function approximation: A gradient boosting machine*, J.H. Friedman, https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full
5. SHAP model overview repository, https://github.com/slundberg/shap
6. Python partial-dependence plot toolbox repository https://github.com/SauceCat/PDPbox
7. *Why should I trust you? Explaining the predictions of any machine learning classifier repository*, by Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, https://github.com/marcotcr/lime
8. *Diverse Counterfactual Explanations (DiCE) for ML*, Ramaravind K. Mothilal, Amit Sharma, Chenhao Tan, https://github.com/interpretml/DiCE

Glossary

Abbreviation	Description
AI	Artificial intelligence
API	Application programming interface
DL	Deep learning
LIME	Local interpretable model-agnostic explanations
ML	Machine learning
PDP	Partial dependence plot
R&D	Research and development
XAI	Explainable AI
EBM	Explainable Boosting Machines

Authors

Ayan Chaki
Head of AI Center of Excellence
Atos AI COE team

Ayan Chaki is a seasoned AI professional with more than 20 years' experience specializing in natural language processing, image processing, computer vision, deep learning, machine learning and audio analytics. He has developed products and solutions for various business domains like banking, financial services and insurance; manufacturing, automotive and aerospace. With multiple granted patents, papers and awards, in his current role he is the global head of the AI Center of Excellence at Atos.

E: ayan.chaki@atos.net
T: +91 (961) 161 3423

Snehashis Panigrahi
Senior AI Practitioner
Atos AI COE team

Snehashis Panigrahi is a senior AI practitioner on the Atos AI COE team and leads a team of AI practitioners in India. He holds a Master of Science in data science from Liverpool John Moores University and more than 18 years of industry experience in various domains and sectors. His experience ranges from team lead to product management to research leadership. Snehashis' s areas of interest lie predominately in responsible AI, explainable AI, natural language processing, computer vision, reinforcement learning, augmented reality and quantum computing.

E: snehashis.panigrahi@atos.net
T: +91 (973) 115 1676

Prakash Kumar Gupta
AI Practitioner
Atos AI COE Team

Prakash Kumar Gupta is an AI practitioner on the Atos AI COE Team focusing on responsible AI and trustworthy AI technologies including explainability, safety and verification in AI systems. He has more than 6 years of industry experience, with interests predominately in responsible AI, explainable AI, machine learning, deep learning and natural language processing.

E: prakash-kumar.gupta@atos.net
T: +91 (869) 956 8735

About Atos

Atos is a global leader in digital transformation with 111,000 employees and annual revenue of c. € 11 billion. European number one in cybersecurity, cloud and high performance computing, the Group provides tailored end-to-end solutions for all industries in 71 countries. A pioneer in decarbonization services and products, Atos is committed to a secure and decarbonized digital for its clients. Atos is an SE (Societas Europaea), listed on Euronext Paris and included in the CAC 40 ESG and Next 20 indexes.

The purpose of Atos is to help design the future of the information space. Its expertise and services support the development of knowledge, education and research in a multicultural approach and contribute to the development of scientific and technological excellence. Across the world, the Group enables its customers and employees, and members of societies at large to live, work and develop sustainably, in a safe and secure information space.

Find out more about us
atos.net
atos.net/career

Let's start a discussion together

CT-220802-EM-WP-Atos-XAI - V6