AWS Whitepaper

Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions

Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions: AWS Whitepaper

Table of Contents

Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions

Publication date: **September 10, 2021** (*Document history*)

Organizations now utilize artificial intelligence and machine learning (AI/ML) solutions to transform their businesses. With this transformation comes the need to ensure that AI/ML models are trustworthy and understandable. This whitepaper outlines the application of model explainability with real-world use cases for institutions using ML. It describes how you can apply model explainability methods to your Amazon Web Services (AWS) AI/ML solutions to meet regulatory compliances, ensure stakeholder trust, provide model transparency, and add business value. This whitepaper is intended for business and technical leaders who are pursuing AI/ML solutions and want additional business value and AI/ML trust by adopting model explainability within their organizations.

Introduction

The purpose of model explainability is to create an understandable solution which can communicate results of AI/ML technology. This field has been expressed as explainable artificial intelligence. Because AI/ML methods have increased in complexity to satisfy industry needs, the requirement for model explainability has risen. When AI/ML solutions are launched into production within customer AWS environments, business leaders or AI/ML owners must trust non-human results that can directly impact business goals.

By using the best model explainability method based on an AI/ML use case, customers can trust an automated solution to meet business objectives. This paper serves as a guide to:

- Understand model explainability and differentiate between interpretability versus explainability given respective applications.
- Utilize a model explainability assessment score card to determine optimal methods and tools to satisfy business requirements.
- Accelerate explainability initiatives by comparing provided common industry use cases.

Are you Well-Architected?

The AWS Well-Architected Framework helps you understand the pros and cons of the decisions you make when building systems in the cloud. The six pillars of the Framework allow you to learn architectural best practices for designing and operating reliable, secure, efficient, cost-effective, and sustainable systems. Using the AWS Well-Architected Tool, available at no charge in the AWS Management Console, you can review your workloads against these best practices by answering a set of questions for each pillar.

In the Machine Learning Lens, we focus on how to design, deploy, and architect your machine learning workloads in the AWS Cloud. This lens adds to the best practices described in the Well-Architected Framework.

For more expert guidance and best practices for your cloud architecture—reference architecture deployments, diagrams, and whitepapers—refer to the AWS Architecture Center.

Interpretability versus explainability

For AI/ML methods, the terms *interpretability* and *explainability* are commonly interchangeable. It is important to distinguish the difference between explainability and interpretability to help organizations determine an AI/ML approach to meet their use case.

Interpretability — If a business wants high model transparency and wants to understand exactly why and how the model is generating predictions, they need to observe the inner mechanics of the AI/ML method. This leads to interpreting the model's weights and features to determine the given output. This is interpretability.

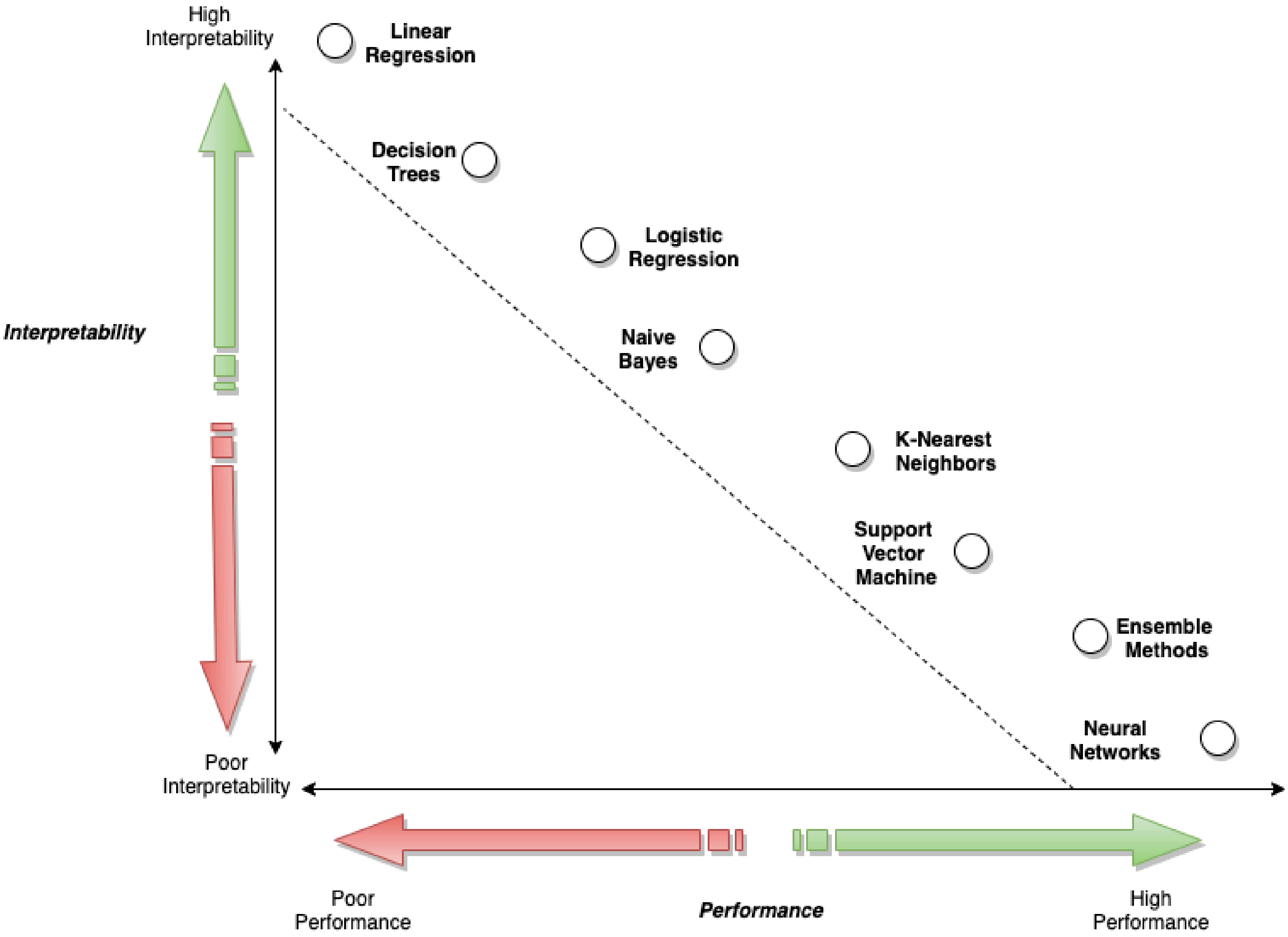
For example, an economist may want to build a multi-variate regression model to predict an inflation rate, they can view the estimated parameters of the model's variables to measure the expected output given different data examples. In this case, full transparency is given and the economist can answer the exact *why* and *how* of the model's behavior.

However, high interpretability typically comes at the cost of performance, as seen in the following figure. If a company wants to achieve high performance but still wants to have a general understanding of the model behavior, model explainability starts to play a larger role.

Explainability — Explainability is how to take an ML model and explain the behavior in human terms. With complex models (for example, black boxes), you cannot fully understand how and why the inner mechanics impact the prediction. However, through model agnostic methods (for example, partial dependence plots, SHapley Additive exPlanations (SHAP) dependence plots, or surrogate models) you can discover meaning between input data attributions and model outputs, which enables you to explain the nature and behavior of the AI/ML model.

For example, a news media outlet uses a neural network to assign categories to different articles. The news outlet cannot interpret the model in depth; however, they can use a model agnostic approach to evaluate the input article data versus the model predictions. With this approach, they find that the model is assigning the *Sports* category to business articles that mention sport organizations. Although the news outlet did not use model interpretability, they were still able to derive an explainable answer to reveal the model's behavior.

When starting a new AI/ML project, you need to consider whether interpretability is required. Model explainability can be used in any AI/ML use case, but if detailed transparency is required, then your AI/ML method selection becomes limited.



Interpretability versus performance trade-off given common ML algorithms

When datasets are large and the data is related to images or text, neural networks can meet the customer's AI/ML objective with high performance. In such cases, where complex methods are required to maximize performance, data scientists may focus on model explainability instead of interpretability.

When starting a new AI/ML project, address interpretability requirements by asking the following questions:

- **Is interpretability a hard business requirement?** If there are regulations or business requirements for complete model transparency, you need to select an interpretable model. This enables you to document how the inner mechanisms of the model impact the output and explain the model in human terms.

- **Can my dataset be used on a simpler model?** Start simple first. If you can meet the objective using an interpretable AI/ML method with full transparency, select that approach. In cases where your data consists of audio, image, or text data types, a more complex AI/ML method may be a better option. Although explaining black box models can be challenging at times, model agnostic methods can provide visibility to derive model explainability in human terms.

A model explainability assessment and communication with AI/ML practitioners (for example, data scientists) can help answer these questions and determine the best model explainability methods to meet your business objective.

Model explainability assessment

As ML projects become complex and vital to solve business objectives, it is critical for business leaders to understand their AI/ML solutions to maximize their return on investment (ROI). To fully use the opportunities model explainability brings to an organization, business leadership should closely collaborate with technical owners to assess the following pillars when pursuing an AI/ML solution.

Topics

Model I/O comprehension

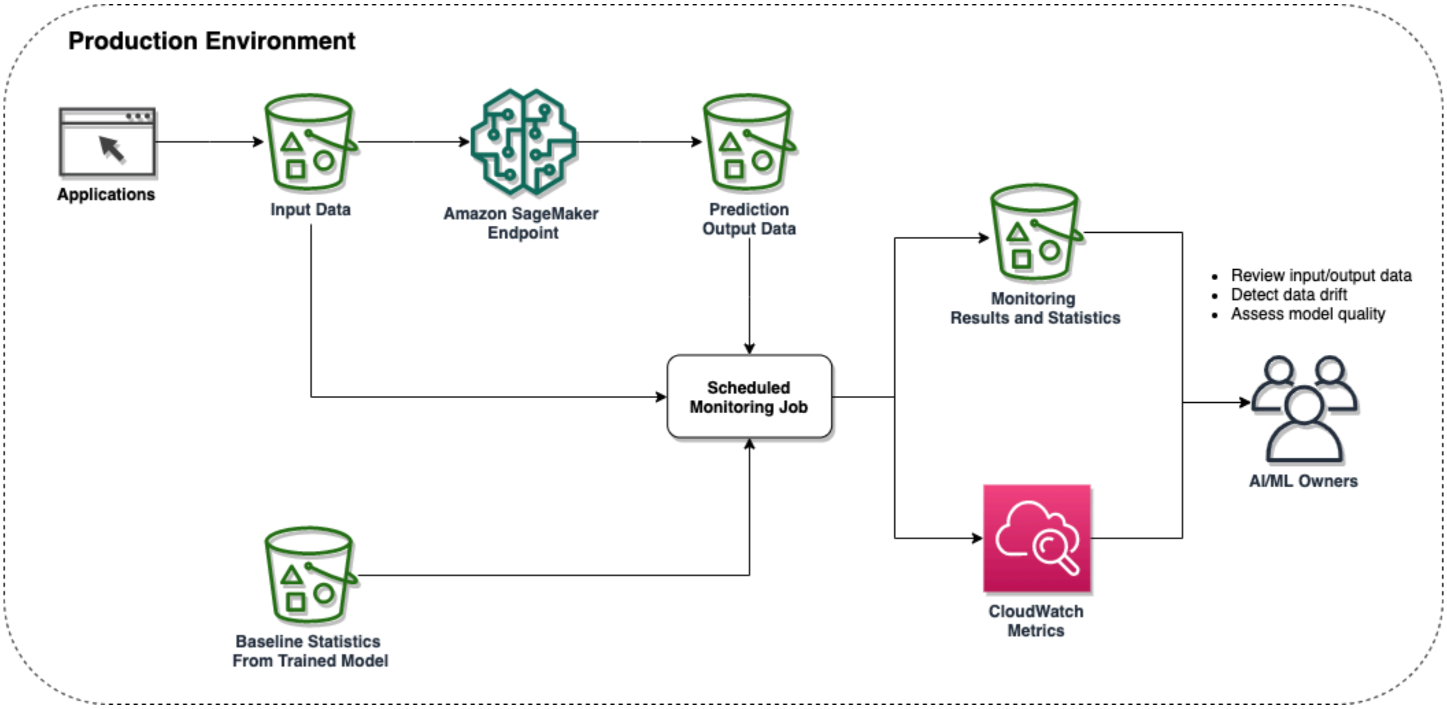
Understanding model I/O means comprehending the semantics of data that is fed into the model, and the output/predictions that are produced by the model. This understanding can be implemented in the development and production stages of the ML lifecycle. Amazon AI/ML services such as Amazon SageMaker Clarify and Amazon SageMaker Model Monitor enable AI/ML experts to detect bias in initial model training while identifying I/O quality when the model is in a production environment. Implementing these I/O comprehension features enables AI/ML owners to understand a model's behavior, and how it treats certain data input samples without requiring advanced knowledge of the model itself.

In the development stage of ML, data scientists evaluate data quality and model performance when training an ML model.

For a model that lacks interpretability, a data scientist can express model behavior by analyzing the input/prediction data and communicating their findings to business owners. To support this model analysis task, services such as SageMaker Clarify detect potential bias during data preparation, after model training, and in deployed models.

For example, a data scientist builds a model to approve finance loans. SageMaker Clarify may notice that a particular attribute, such as occupation, is impacting the predictions, and is causing a bias within the model. This enables a business owner to understand how the model consumes the data, and can create an opportunity to capture domain knowledge to correct the bias. Because domain knowledge often helps model development and performance, business leaders can contribute their expertise if the model is not utilizing specific information. This is where biases can be eliminated with the help of either removing biased data or providing key information for data model inputs.

In the production stage of ML, the model processes new input data and produces predictions continuously. In this stage, a data scientist is not available to explain the inner mechanics of the model or communicate the model's behavior. Rather, the I/O are used to explain the behavior of the model. SageMaker Model Monitor provides a way to continuously monitor the quality of an ML model in production while enabling teams to utilize prebuilt monitoring rules, or by creating custom code to evaluate data inputs and predictions.



Model monitoring in production environment

As shown in the preceding figure, SageMaker Model Monitor uses data from artifacts created when training the production model and comparing the I/O of a live SageMaker Endpoint. AI/ML owners can continuously monitor the data or create events to assess both data and model quality.

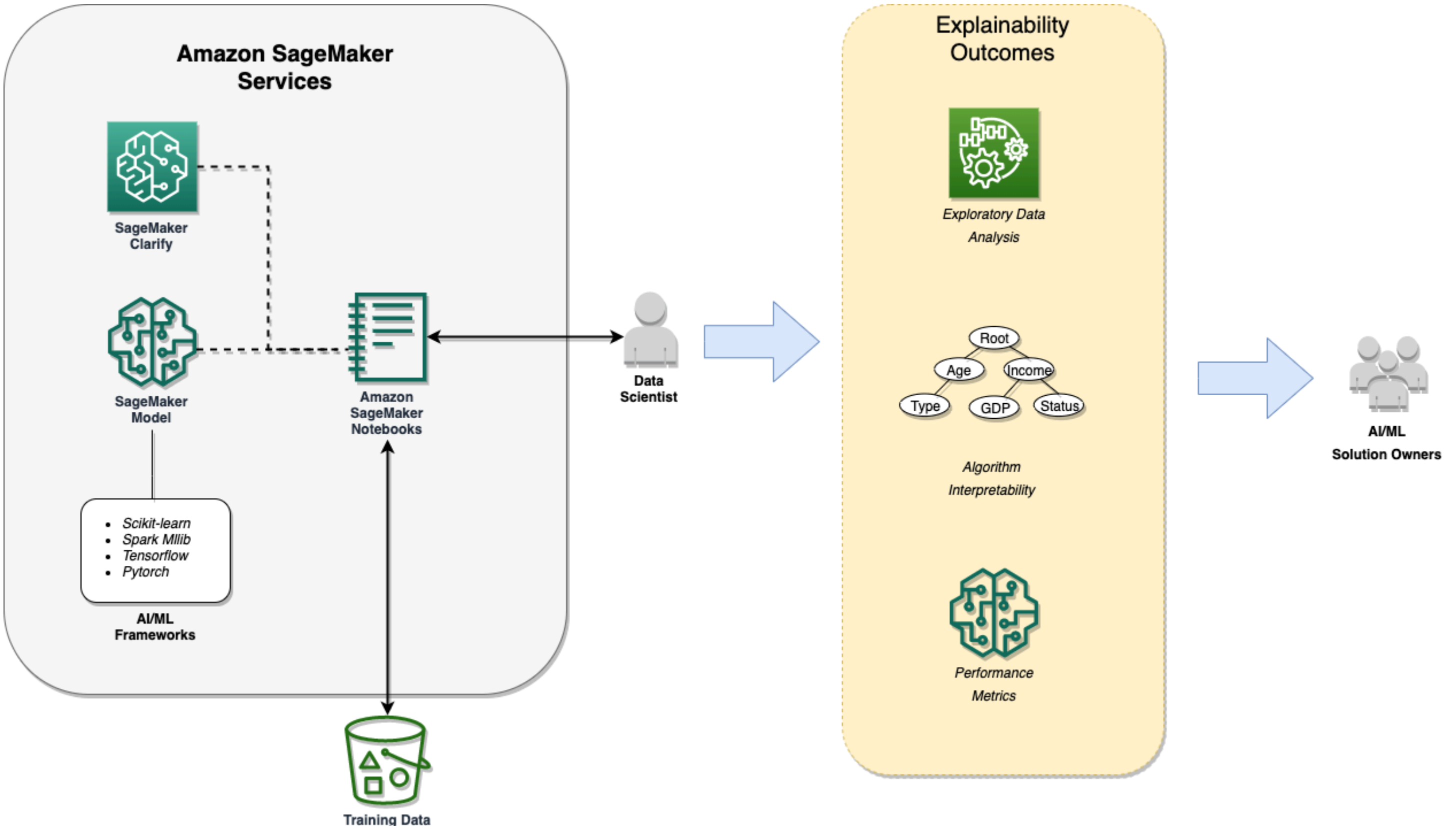
With this assessment, teams can derive and explain the model's behavior. Additionally, SageMaker Clarify can fit into this process to check whether a model starts to include biasness in predictions. For example, when predicting financial loan approvals, predictions can start to favor one demographic over another, such as high-income earners receiving more positive predictions compared to low-income earners as data input changes.

Model transparency

ML owners require a degree of model transparency to match the AI/ML method to the business objective. This is either a high level understanding to conceptualize the impact of the ML model in real world environments, or a deep level comprising how the method functions internally. The level of model transparency depends on the knowledge required to understand the internal mechanics of the ML algorithm. In initial stages of AI/ML development, you should consider trade-offs between interpretability versus performance with regulatory requirements in mind.

Occasionally, regulators require evidence to justify how the ML model works. In these scenarios, the technical practitioner must present how the ML model functions with related data and provide artifacts as evidence. These artifacts can be used to explain positive or negative model impacts to real-world processes.

If regulatory requirements are present, it is recommended to initially investigate approaches utilizing interpretable algorithms. Data scientists can utilize AI/ML frameworks such as scikit-learn with SageMaker, and build an interpretable AI/ML model. With SageMaker Notebooks, AI/ML practitioners can document each model building step from exploratory data analysis to model building, and then to model deployment.



Using SageMaker services to communicate model explainability

Additionally, the AI/ML practitioner can communicate specific model parameters (for example, decision tree path given decision tree model) and show the method through a SageMaker Notebook. The notebook can be saved and pushed to an internal repository, which is archived and shared with regulation teams or AI/ML owners, as depicted in the preceding figure.

When interpretable model performance cannot meet business objectives but model transparency is required, you need to either visit other pillars mentioned in this section, or pursue a model agnostic approach. To pursue a high-level understanding of the AI/ML model, business owners should ask AI/ML practitioners to use model agnostic approaches to answer common real-world questions such as:

- Why did this email get flagged as spam?
- How did this person’s loan application get rejected?
- What data features are causing the model to recommend these product types?

These types of questions can be answered by using model agnostic approaches that include methods such as feature attribution, local interpretable explanations (LIME), and use of surrogate models.

On AWS, AI/ML practitioners can use Amazon Sagemaker Clarify, which uses Shapley values to help answer how different variables influence model behavior. These techniques help derive explainability to help business leaders reach a level of model transparency to understand and meet their business goals.

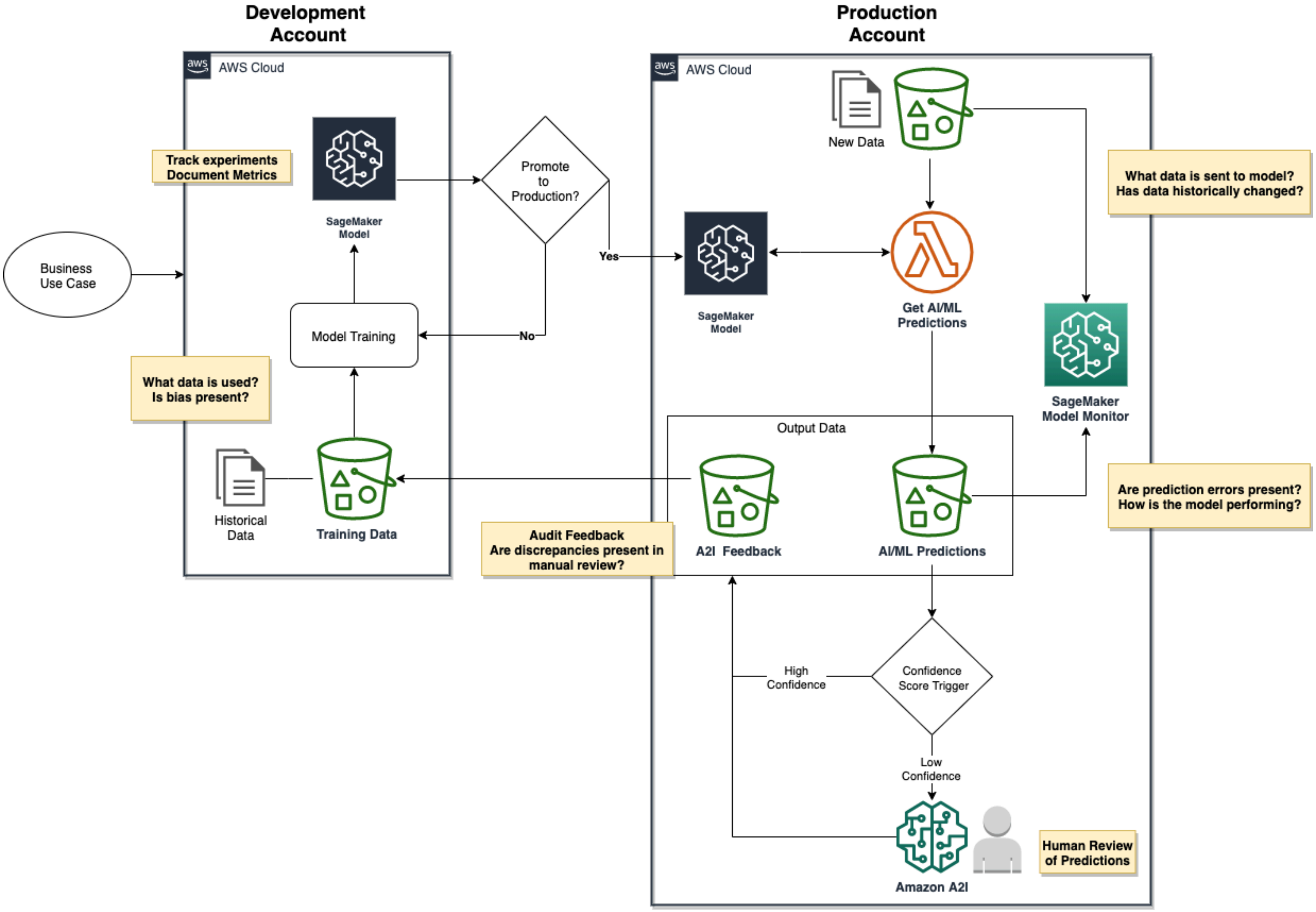
Although it is not required for business leaders to fully understand the ML algorithm, having high-level knowledge of how the model behaves with given data can help business leaders conceptualize the model's implementation. This provides business leaders with context and an intuitive understanding behind any model shortcomings. To achieve this, business leaders should ask the technical practitioners to explain the model in human terms related to the context of the addressed business objectives.

Model confidence

Gaining model confidence means business leaders have trust in model results. To achieve this, models should produce consistent performance and prediction output should be audited. As discussed earlier, SageMaker Model Monitor provides a method of monitoring that helps track model performance over time. With this comes the ability to track input data and prediction output distributions for model drift.

Another way to gain human confidence in an AI/ML solution is through the use of confidence scores with predicted output. Technical practitioners should store confidence scores in an AWS data storage service such as Amazon Simple Storage Service (Amazon S3) or Amazon DynamoDB to alert AI/ML owners or application users of model degradation.

Given an active model, model predictions can be reviewed by confidence scores to help review and audit model performance. Domain experts can either review predicted data with poor confidence scores in real time or periodically review random samples of predictions for auditing purposes.



AI/ML lifecycle with Amazon A2I and SageMaker Model Monitor

Amazon Augmented AI (Amazon A2I) provides built-in human review workflows for multiple ML use cases. As shown in the preceding figure, Amazon A2I enables domain experts to review predictions from a SageMaker model. Poor predictions with low confidence scores invoke a human feedback loop to allow humans to verify or supply the true answer to the predicted output. This feedback is reviewed by AI/ML technicians or business owners to evaluate the model output versus the data's ground truth.

Amazon A2I human review workflows integrate with other AI/ML services such as Amazon Comprehend, Amazon Textract, and Amazon Rekognition. By using these AWS AI/ML services to monitor models and initiate human feedback loops, leaders can utilize this audit information to gain confidence in model performance and allow model inference in a live application.

Business value creation

Enabling model explainability creates business value. After business leaders understand the models and trust the model output, the next step is to assess how actionable the model outputs are to drive business impact. For example, in a pricing optimization use case, if the model can output the change of a price and its impact on bottom line and probability, then the outcome becomes more actionable.

Business leaders can use model output information to make data-driven decisions to meet their organization's goals. For example, Amazon Forecast, a service that uses time series data to build forecasts, offers quantiles at which probabilistic forecasts are generated. Using these quantile ranges enable business users to understand forecast options while simplifying decision processes.

For another example, for a product marketing campaign business case, the recommendation model may have a bias and only output popular products. Given that business owners can understand this bias, they can delay releasing the AI/ML solution publicly to avoid a decrease in business value. Amazon Personalize, a service that uses data to build recommendation models for personalized real-time recommendations, has features such as exploration weight and age cut off that can be used to explore such opportunities.

Model explainability assessment score card

Before initiating an AI/ML project, business leaders can work with technical counterparts to determine explainability requirements by assessing the following pillars of model explainability:

- Model I/O comprehension
- Model transparency
- Model confidence
- Business value creation

The following score card assesses each pillar, and serves as a guide to evaluate the need for model explainability and help business owners decide how they want to understand the technical solution. For each assessment section, if the sum of scores is greater than ten, it is recommended that businesses leaders dive deep and explore that pillar. This whitepaper discusses each assessment section previously, giving the reader the ability to initiate exploration of the four provided pillars of model explainability.

A total sum of scores is used as the aspirational goal for the project. After a project is finished, a post assessment is conducted to identity gaps, and evaluate if the goal was achieved given the users' answers.

Table 1 – Model explainability assessment score card

Assessment sections	Score (0-5)
Model I/O comprehension 1. Do I need to understand the model input data or the model predictions in this use case? Do I care if the data has a bias?	
1.1. I require a general idea of the input data and how the model uses it. I need to understand what the model outputs mean.	
1.2. I require a full understanding of how the model uses the input data. I need to fully understand the model output predictions. I must validate all model outputs with domain knowledge.	
Model transparency 2. Do I need to understand how the model works in my use case?	
2.1. I need a general idea of how the model works, not the inner mechanics of the model.	
2.2. Full transparency of the model is required. Documented inner mechanics of the model is required. The model must be presented in human terms as much as possible.	
Model confidence	

Assessment sections	Score (0-5)
3. Do I need to trust the model outputs in my use case?	
3.1. I require trust in the model’s ability to provide the right predictions the majority of the time. I can occasionally accept wrong predictions without explanation.	
3.2. I require trust in the model to consistently provide correct predictions. When predictions are incorrect, I require review and knowledge of why it was incorrect.	
Business value creation 4. Do I need to understand how the model output can drive business impact?	
4.1. I require an understanding of how the model can drive business impact. It is acceptable if partial information from the model output is not actionable.	
4.2. I must fully incorporate learnings from the model to drive impact; the model provides actionable information to generate business value.	

Score

5 – Very important **4** – Important **3** – Neutral **2** – Not very important **1**–Not at all important

After completing the assessment, the team or business owners may have a prioritized pillar. For the prioritized pillar, it is recommended they understand common technical constraints and investigate applicable AI/ML services. The following table complements this recommendation and provides common algorithm restrictions and AI services per pillar to help accelerate the adoption of model explainability and increase project success.

Table 2 – Common method restrictions and tools per pillar

Pillar	Algorithm restrictions	Tools and AI services
Model I/O comprehension	No restrictions on algorithm s (choose the algorithm with the highest performance)	• SageMaker Clarify • SageMaker Model Monitor
Model transparency	Highly interpretable techniques such as linear regressions or tree-based methods	• Scikit-learn with Amazon SageMaker
Model confidence	Stable algorithms that can output usable confidence scores	• Amazon A2l • SageMaker Model Monitor
Business value creation	No restrictions on algorithm s, but the algorithms should be flexible enough to support different business use cases and scenario planning	• Forecast* • Amazon Personalize*

* Forecast and Amazon Personalize support specific use cases.

For example, if an AI/ML team decides that model transparency is the most important pillar to satisfy the use case, the technical counterparts should consider implementing highly interpretable ML algorithms. SageMaker currently offers 17 built-in algorithms and several toolkits that can aid in this effort.

Getting started: Common business use cases

Common business use cases involve either explaining a proposed ML model to support customer decisions, or using explainability to meet security or auditing compliances. Business use cases across industries can involve the following requirements:

- Business leaders require general knowledge and understanding of the delivered AI/ML solution. Leaders expect to visualize and seek the *why* and *how* related to the technology.
- Business requires evidence to understand and trust an automated solution before replacing previous processes (for example, human processes and rule-based methods).
- Security or auditing team require evidence to validate the AI/ML predictions. Predictions can have a direct impact on business performance. Compliance teams want to reference the root cause of AI/ML impacts given good or bad outcomes.

It is recommended to address these common requirements before initiating the AI/ML solution and lifecycle. If strict guidelines for model interpretability is a requirement, your first action should be trying to find the best solution to meet model transparency. This will mitigate the need to redevelop an AI/ML solution if a model's complexity deters its usage.

Common industries

With common business use cases and the requirement to trust AI/ML solutions, multiple industries face challenges in using model explainability. Primary industries that use model explainability are:

- Financial services
- Health care
- Manufacturing
- Energy
- Other (education, sports, and so on)

The following examples illustrate scenarios with solutions of model explainability across industries where knowledge and model trust is key.

To help associate domains of explainability with real-world AI/ML scenarios, each industry use case is tagged with one or more of the model explainability pillars.

Financial services

With the rise of automatic investment accounts in North America, and an increasing number of online footprints, financial services has become one of the core industries to adopt ML and reap its benefits. As the industry matures, fairness, transparency, and actionability are the three pillars critical to ensure success.

Loan approval process

Pillars: Model I/O comprehension, Model transparency

One of the most popular use cases in ML is automating loan approval processes and calculating credit scores. Historically, lenders relied on specialized personnel to review consumer credit reports to decide loan approval. This manual approach leads to the increase in human errors and default risks.

With the aid of ML, companies utilize additional support data to lend money to customers that are statistically likely to pay the lender back, they can also lean out labor-intensive tasks through process automation. To ensure that ML models do not make the same mistakes as humans, model explainability is essential to verify that the features the models rely on to make decisions are not prone to human biases. Having model explainability helps companies decide how they improve their application process by identifying key elements the model predicts as important.

Pricing and forecasting

Pillar: Business value creation

Time series forecasting methods have gained popularity in predicting supply, demand, and inventory on the market. Enabling model explainability helps professionals understand variables the model deems as important behind prices and market. Seasoned professionals use this information with their own expertise to achieve informed and accurate decisions on investment, thus increasing their ROI. Models can summarize useful information that will enable businesses to react faster to the market. Without having to spend time trying to gather this information manually, human resources are now directed to areas of the business where they add greater value.

Fraud detection

Pillar: Model confidence

Another popular use case in the financial industry is fraud detection where ML models are used to detect fraudulent activities by identifying anomalies. This use case helps businesses save money and recover losses. In fraud detection, domain expertise typically helps when reviewing fraud. When a trained model outputs a confidence score to indicate fraud, the prediction can be sent to specialized personnel who can validate the decision. In these cases, a model’s ability to explain its prediction reasoning is critical to determine fraud.

Health care

ML is gaining traction in the health care industry, from medical imaging and diagnostics to revolutionizing decision support systems. Because model predictions can have far reaching consequences and a mistake may cost lives, explanation is required to ensure model consistency and validity. From a legal perspective, the FDA requires all medical recommendations to be explainable.

Medical imaging and diagnostics

Pillar: Model confidence

Computer vision techniques are popular in helping physicians formulate diagnoses. Because the industry matures, AI/ML approaches have the potential to advance the quality of patient outcomes. The consequences of a wrong diagnosis are dire. It is critical for doctors to validate the results accurately. Implementing model explainability methods help provide context to doctors and enable justification in clinical decision making or verifying a suspicious diagnosis. With models providing transparency, clinicians gain trust in the automated solution and adopt AI/ML into their daily workflow, providing clinicians time to focus on prioritized endeavors.

Admission prediction

Pillar: Model I/O comprehension

ML is used to predict the likelihood of patients being admitted to the hospital from the emergency room. Instead of having physicians manually review each case, physicians can use AI/ML predictions to review key factors associated with patients.

With explainable prediction data, physicians are enabled to reach informed decisions that improve hospital efficiency and patient satisfaction.

Manufacturing

In manufacturing it is common for automation to replace historical production processes. It is vital for the customer to understand how and why the AI/ML solution performs to justify the manual replacement with automation.

Device maintenance and repair

Pillar: Model confidence

Technicians have been known to use domain knowledge when detecting failing factory devices. An AI/ML solution can measure quantitative data (for example, temperature, age, and throughput) for each of the devices and use this to predict a future time window in which the device is likely to fail. Every device replacement accrues cost and time, and technicians can be hesitant to trust device failure predictions. To mitigate this concern, feature attribution can be used for the technician to understand which metrics are impacting a model's prediction. Given this knowledge, model explainability justifies the decision to replace the failing device.

Factory machine automation

Pillars: Model confidence, Model transparency

Production processes are being replaced and automated with machines or AI/ML methods. In these situations, quality output is a significant requirement. A computer vision AI/ML model detects weaknesses or quality issues by scanning products in an assembly line. A computer vision model is commonly perceived as a black-box solution. Given the model's complexities, a monitoring system will allow transparency of the solution. A monitoring system evaluates data inputs and outputs of the computer vision model in real time. This process enables auditing controls which mitigates product quality issues.

Energy

As the energy industry expands to alternative energy sources and automated technologies, AI/ML is becoming increasingly relevant in energy distribution. Because energy can impact thousands to millions of people, it is essential to understand and gain trust in utilizing automated solutions to explain energy output.

Smart grid technology

Pillar: Model transparency

Smart grids intelligently support the generation, transmission, and distribution of electrical power. An AI/ML solution provides automated decisions to distribute electricity. This decision is not taken lightly because businesses and homeowners are impacted by the supply of energy. In these scenarios, methods are established to capture why the AI/ML solution is impacting energy grid outcomes. Algorithms, such as decision trees, provide a decision path which explains the model output. With this information, energy providers are empowered to provide suitable reasons of power distribution to their customers.

Power grid stabilization

Pillars: Model transparency, Model I/O comprehension

The ability to proactively stabilize power grids mitigates the risk of blackouts affecting thousands of people. An AI/ML model is used to detect anomalies to help predict power outages. Although it is important to be alerted to possible failures, root cause information is required to help engineers determine mitigation efforts. In this case, engineers assess the feature importance and capture input data when addressing the anomaly. This identifies the root cause of the issue, provides explanations, and mitigates a power outage.

Public sector

Automated technology is used in society to support governmental services such as military, law enforcement, and public infrastructure. The usage of AI/ML may impact everyday citizens. Explanations are required to ensure automated decisions are not causing negative impacts on human rights and public services.

Traffic monitoring with law enforcement

Pillars: Model transparency, Model I/O comprehension

Law enforcement uses object detection to identify traffic violations. If drivers speed or commit a traffic violation, a camera uses an AI/ML model to capture the image and determine the offense. In these scenarios, drivers desire an explanation and evidence of the violation. The quality of the technology can come into question. Proactive model monitoring controls help detect biases and performance issues to gain public trust.

Along with monitoring, a feature attribution approach is used to explain how the computer vision model is interpreting the image or series of images. This method supports law enforcement to provide evidence and justification for traffic violations.

Policy support

Pillar: Model transparency

Demographics and geographic public data types are used by government departments, such as transportation, to help support policy decisions. An AI/ML model is utilized to capture traffic and economic data to prioritize road maintenance and repairs. Planning resources use the model predictions to make planning decisions that will impact communities. In these scenarios, policy leaders need to ensure a well-informed decision. This requires understanding the main contributing features of the AI/ML prediction and how the model interprets the training data. An interpretive model, such as multi-variate or logistic regression, helps decision makers understand and explain how each variable of the dataset impacts their final decision.

Other industries

Autonomous vehicles

Pillars: Model confidence, Model I/O comprehension

Autonomous vehicles make decisions from real-time image processing. Such decisions lead to life and death scenarios. By understanding the training data set and how the AI/ML solution determines a decision enables manufacturers to explain the ethical stance their company makes in life-changing scenarios.

Education

Pillars: Model confidence, Model I/O comprehension

In the education sector, educators use AI/ML methods to recommend material for curriculum planning. To deter bias and mitigate ethical concerns, human understandable evidence must be involved with a model's recommendation.

Media and entertainment

Pillar: Business value creation

In media and entertainment, recommendation systems are used to provide content to users (for example, Netflix, HBO, and YouTube). Marketing campaigns desire the knowledge of why and how viewers are receiving content by automated methods.

Conclusion

This whitepaper described how model explainability relates to AI/ML solutions, giving customers insight to explainability requirements when initiating AI/ML use cases. Using AWS, four pillars were presented to assess model explainability options to bridge knowledge gaps and requirements for simple to complex algorithms. To help convey how these model explainability options relate to real-world scenarios, examples from a range of industries were demonstrated. It is recommended that AI/ML owners or business leaders follow these steps when initiating a new AI/ML solution:

- Collect business requirements to identify the level of explainability required for your business to accept the solution.
- Based on business requirements, implement an assessment for model explainability.
- Work with an AI/ML technician to communicate model explainability assessment and find the optimal AI/ML solution to meet your business objectives.
- After the solution is completed, revisit the model explainability assessment to evaluate that business requirements are continuously met.

By taking these steps, you will mitigate regulation risks and ensure trust in your model. With this trust, when the event comes to push your AI/ML solution into an AWS production environment, you will be ready to create business value for your use case.

Contributors

Contributors to this document include:

- Joe King, Data Scientist, AWS Professional Services
- Betty Zhang, Data Scientist, AWS Professional Services
- Hanif Mahboobi, Senior Data Science Manager, AWS Professional Services
- Shantu Roy, Global Director of AI/ML, HPC, Quantum Computing, AWS Professional Services

Further reading

For additional information related to model explainability and AWS, see:

- Amazon AI Fairness and Explainability Whitepaper
- Fairness Measure for Machine Learning in Finance
- Interpretable Machine Learning – A Guide for Making Black Box Models Explainable
- ML model explainability with Amazon SageMaker Clarify and the SKLearn pre-built container
- AWS Architecture Center

Document history

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Initial publication	Whitepaper first published.	September 10, 2021

Note

To subscribe to RSS updates, you must have an RSS plug-in enabled for the browser that you are using.

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.