INSTITUTE OF INTERNATIONAL FINANCE
OliverWyman Forum

Briefing Note | **DataTalk**

03

Artificial Intelligence / Machine Learning Explainability

June 2021

Financial institutions (FIs) are increasingly employing these technologies to gain competitive advantage, better leverage data and increase efficiencies. Organizations need to be able to explain how their AI and ML models work to build trust with customers and regulators. Leading industry practitioners discussed the issues around explainability in the third session of DataTalk. This note provides a brief summary of the key themes from the discussion, respecting that the conversation was conducted under the Chatham House rule, and comments are not attributed.

Explainability is a trust issue. Regulators are concerned with levels of complexity that offer limited or no insight into how the model works (i.e. so called black boxes). They want to know how models, including AI and ML, reach decisions on extending or denying credit, whether FIs have appropriate risk controls in place, and the like. A challenge of using AI/ML models is often the lack of transparency, which is imperative to building trust with customers and other stakeholders. Banks have long had to explain their decision processes to improve confidence in the robustness of a model. Demystifying the concept can help institutions understand the different types of explainability tools available and their implications.

Start with high-level principles. Many begin by developing their own foundational principles or build on outside ones, such as the Monetary Authority of Singapore's principles to promote Fairness, Ethics, Accountability and Transparency, or FEAT. In some cases, firms then broaden their approach to develop a company-wide data ethics framework that extends beyond AI. Banks can show they are living up to their values by promising that AI will be used fairly; that they will be transparent about how data is collected, safely stored, and used.

Taking a layered approach. Many firms start by assessing the risks of how material a particular AI/ML use case is - materiality is assessed across several dimensions, from the impact on people to the financial impact on the bank to regulatory impact. The greater the materiality, the greater the risk and the greater the need for explanations of the decision outcomes. This is especially the case for decisions affecting people, such as credit scoring.

Inherent vs. Post-hoc explainability. FIs are using a variety of techniques to explain their AI/ML decision processes. For high-risk applications, many prefer ML models that are inherently explainable, which allow developers and other stakeholders to understand how they make predictions. For other applications, post-hoc explainability (i.e., deriving explanations after the training) may suffice. FIs may have different views on which technique suits particular use cases, but when post-hoc techniques are used firms should implement appropriate controls.

Context matters. Explainability can be just as important for internal purposes. It can serve as a diagnostic tool to mitigate bias in lending and other outcomes. ML can also provide opportunities to make corrections when unfair bias is identified, and to better organize and understand data, in ways that were not possible with traditional analytics. A governance framework with tools and processes to test, monitor, and govern ML models is key. Firms also may want to ask themselves whether they are holding machines to a higher standard that human decision-makers.

Regulation vs. supervision. Regulation should be technology-neutral. Regulators want FIs to have the appropriate knowledge, sophistication, and controls for whatever technology they are using. They do not want to hard-code regulation to specific tech tools that could quickly become outdated. This implies that better supervision may be more effective than more regulation. And it would be helpful to the industry if supervisors would make clear what AI/ML use cases they want to prioritize.