INSTITUTE OF
INTERNATIONAL
FINANCE

June 30, 2021

Board of Governors of the Federal Reserve System
20th Street and Constitution Avenue NW
Washington, DC 20551
Docket No. OP–1743.

Via email: regs.comments@federalreserve.gov

Re: Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning

The Institute of International Finance (IIF) welcomes the interagency Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning, published on March 31, 2021 by the Board of Governors of the Federal Reserve System, the Consumer Financial Protection Bureau, the Federal Deposit Insurance Corporation, the National Credit Union Administration, and the Office of the Comptroller of the Currency (Agencies).

We appreciate the opportunity to comment and are supportive of regulatory efforts to promote responsible adoption of artificial intelligence (AI) and machine learning (ML). Financial institutions (FIs) share the view that the use of AI and ML should lead to fair and ethical outcomes, and uphold principles such as accountability, fairness, and transparency. We view the strong history of collaboration between policymakers and industry as a fundamental building block towards a sound and responsible approach to developing policy for AI and ML.

In our response, we provide data points and examples drawn from our previous IIF surveys on the use of machine learning in credit risk, anti-money laundering and financial crime prevention, and around the end-to-end governance of machine learning models.[1] Additionally, we drew examples from our Machine Learning Thematic Series papers, which covered explainability, bias and ethics, as well as provided recommendations to policymakers. Our response focuses mainly on FIs' use of ML, because of the depth of data we have gathered through our surveys. Given that the research continues to develop in the area of AI and ML, it is our view that regulatory initiatives should remain dynamic, technology-neutral, and futureproof. It is of upmost importance for supervisors and regulators to continue to encourage innovation and show tolerance as financial institutions use AI and ML.

Similarly, we share the view that regulatory initiatives should take a risk-based approach to determine appropriate controls that are commensurate with the risk of each specific use case.

[1] In March 2018, the IIF published its Machine Learning in Credit Risk Report (ML-CR Report), in which we surveyed a globally diverse sample of 60 firms on their applications, motivations, experiences, and challenges as they apply ML techniques in credit risk. In July 2019 we published our Machine Learning in Credit Risk, examining the continuing evolution and progress over the last year and a half. Beyond the IIF credit risk surveys, in October 2018, we published the Machine Learning in Anti Money Laundering Report. This report surveyed 59 firms, the majority of which were also interviewed for the two ML-CR reports. In December 2020, we published the Machine Learning Governance report looking at the end-to-end governance of machine learning models. This report surveyed 66 firms across our membership.

In 2020 we conducted a survey that provided an overview of the governance of ML models – looking at the end-to-end model governance process for ML models. ML is increasingly being used in areas such as credit risk, compliance, market risk assessment and in insurance underwriting, in addition to its use in customer facing areas (such as marketing, and customer service). Most firms in our sample are using ML in production (68%), and over a quarter of participants have active pilot projects in place (26%).

Similarly, our 2019 Machine Learning in Credit Risk study found that the adoption of ML in credit risk modeling and management had nearly doubled when compared to the results of our 2018 study. [2] The sophistication of ML models and the breadth of application across customer segments also saw a significant increase. In the credit risk area, most FIs were using ML focusing on existing retail (consumer) portfolios; where typically FIs possess larger volumes of standardized, high-quality data. However, our 2019 survey results show that ML is increasingly being used for SMEs and other non-retail (i.e., CRE and public sector) portfolios. A recent Bank of England paper also signaled similar results, with the use of ML nearly doubling in the last year.[3]

Our studies show that the adoption of ML techniques continues to deliver tangible benefits to FIs, including improved model accuracy, the ability to overcome data deficiencies and inconsistencies, and discovery of new risk segments or patterns.

For instance, ML's increased analytical power has been used for model development, for model building and variable selection, allowing FIs to filter through several more variables in search of significant predictors. For many firms, the use of ML in model development has resulted in an increase in model accuracy. ML methods enable FIs to develop a multitude of models, number of targets, and multitude of design constructs that allow FIs to see more detailed patterns for segments of the population and gain a greater granular understanding of those patterns. Similarly, ML has helped firms expand services to new customer segments thanks to new data insights. Between 2018 and 2019, the number of FIs deploying models in production for SME portfolios rose by more than 350%.

Although the power of ML in making predictions is often discussed, what is truly impactful is the power to help organize and understand data. In fact, FIs that use ML for this purpose report achieving a faster return of investment, as better data can enable banks to screen customers and transactions more effectively against sanctions lists for instance. In the area of financial crime prevention, ML algorithms are being used to partially automate financial crime investigations.[4]

However, ML works by ingesting historical data and acting on the objectives detailed by the developer. Historical data is inevitably reflective of human biases, including unconscious ones,

[2] IIF, *Machine Learning in Credit Risk*, March 2018; and IIF, *Machine Learning in Credit Risk, 2nd Edition Detailed Report*, July 2019. The full Detailed Reports are limited to official sector and participating firms. The survey results include a wider scope than credit scoring and decisioning, including credit monitoring (including early warning systems), and for collections, restructuring, and recovering. While there are indeed instances where machine learning techniques are being used or explored for modeling purposes, survey results extend to areas that are better characterized as 'Credit Risk Management'. A short-form Summary Report can be accessed at: https://www.iif.com/Publications/ID/3525/Machine-Learning-in-Credit-Risk-2nd-Edition-Summary-Report and https://www.iif.com/publication/regulatory-report/machine-learning-credit-risk

[3] Bank of England, *The Impact of COVID on Machine Learning and Data Science in UK Banking*, December 2020. Can be accessed at: https://www.bankofengland.co.uk/quarterly-bulletin/2020/2020-q4/the-impact-of-covid-on-machine-learning-and-data-science-in-uk-banking

[4] IIF, *Machine Learning in Anti-Money Laundering*, October 2018. The full Detailed Report is limited to official sector and participating firms. A short-form Summary Report can be accessed at: https://www.iif.com/Publications/ID/1421/Machine-Learning-in-Anti-Money-Laundering

and the main concern has been that misguided correlations could have powerful implications given the automated nature of ML algorithms. In particular, that biases inherent in training data can result in models that amplify the biases already present in society.

U.S. FIs have extensive experience in and strong processes in place for managing model risk. The interagency guidance on Model Risk Management (MRM Guidance) has served banks well in managing risks, including models using AI and ML.[5] In fact, many recent proposals on responsible use of AI from across the globe appear to have been influenced by the MRM Guidance. Given the potential for ML to provide a broad range of benefits, any policy action should not constrain the responsible development and innovative use of this technology. In our view, it is important to hold banks and non-banks to similar standard of activities that may result in customer harm.

We highlight that the use of ML largely falls under FIs existing risk management frameworks and risk control measures. While ML models may at times, depending on the use case, present differing risk than traditional models, this is being assessed and adapted to ensure that new risks and responsibilities related to the use of new technologies are considered. We are of the view that all models should be subject to an appropriate control framework, i.e., that FIs should manage the risks of deploying ML by applying governance that is structured to ensure that appropriate controls are in place commensurate with the materiality of each specific use case, regardless of whether AI or ML techniques are used.

In that vein, consumers need equal and consistent protections, regardless of where they secure credit, and work towards financial security. Having consistent model risk management standards among all financial regulators is crucial. A lack of consistency between regulators and institutions leaves customers at risk, especially to those firms outside the well-regulated banking industry.

FIs stand ready to partner with the official sector to come up with more collaborative ways to monitor and manage risks arising from the use of AI and ML. The IIF would be pleased to convene a roundtable to facilitate the dialogue between the industry and policy makers on issues around the use of AI/ML, and we also encourage policymakers to join the discussions with industry experts in our regular DataTalk forum.

In the following Appendices, we address questions on a thematic basis, rather than repeating each theme under each question (please see Appendix A and B).

The IIF looks forward to continuing the dialogue on this important topic, and to contributing to the further development of safe and effective innovations that can benefit the economy and support stability. My colleague Natalia Bailey (nbailey@iif.com) and I (bcarr@iif.com) stand ready to engage in additional discussions and consultations.

Yours sincerely,

Brad Carr

[5] Model Risk Management (MRM) Guidance [Federal Reserve Board Supervisory Letter SR 11-7 (Apr. 4, 2011), OCC Bulletin 2011-12 (Apr. 4, 2011), and FDIC FIL 22-2017 (June 7, 2017)] requires critical analysis throughout the development, implementation, and use of complex algorithms like AI, and sets supervisory expectations for independent review of models to confirm they are functioning as intended.

Managing Director, Digital Finance

Appendix A

In the following, we address the RFI's questions on a thematic basis, rather than repeating each theme under each question.

Explainability:

Question 1: How do financial institutions identify and manage risks relating to AI explainability? What barriers or challenges for explainability exist for developing, adopting, and managing AI?

Question 2: How do financial institutions use post-hoc methods to assist in evaluating conceptual soundness? How common are these methods? Are there limitations of these methods (whether to explain an AI approach's overall operation or to explain a specific prediction or categorization)? If so, please provide details on such limitations.

Question 3: For which uses of AI is lack of explainability more of a challenge? Please describe those challenges in detail. How do financial institutions account for and manage the varied challenges and risks posed by different uses?

A. Machine Learning Governance

Firstly, we agree with the premise that machine learning (ML) models should follow the same strict requirements that are placed on any other type of model. This includes having (i) a robust model development, implementation, and use control framework, (ii) a sound model validation process to ensure that models are performing as expected, and (iii) an effective framework of governance.

Our research indicates that FIs use of ML in a production environment is embedded in the firms' model risk management framework. Our latest study[6] found that 36% of participating firms apply their existing model risk management framework to ML applications, i.e., their framework applied for all types of models for the model life cycle purpose, including ML models. However, our results also show that 44% were moving towards an enhancement of their current model risk management framework to include specific ML considerations to it (15% had developed such enhancement and 29% were in the process of developing one). Typically, such enhancements included expanding the independent model risk and model validation methods and skillsets to better capture ML applications.

At the regional level, our results show that American firms were using their existing MRM framework for ML applications at a noticeably higher level than in other parts of the world. In fact, 55.5% of U.S. firms currently utilize their existing model risk management framework, and about 33% developed an enhancement to their model risk management framework to provide internal guidelines and additional requirements for ML model (e.g., on documentation, training, etc.).

Our survey indicates that the strong processes that U.S. FIs have in place for managing model risk set out by the interagency MRM Guidance, also apply for ML models. In the U.S., ML models are covered under the broader definition of a model as defined in the MRM Guidance. Thus, ML models are risk rated, with more material and complex models being subject to more robust validation and audit requirements. In many cases, methodology is a consideration, which includes

[6] There are nine U.S. firms in our IIF Machine Learning Governance study. "Firms" represent banks and insurers. Firms are categorized by region according to where they are headquartered, while acknowledging that many have operations across multiple jurisdictions.

the technique used, characteristics of the model, volume of data used, degree of transparency, and in some cases out-of-sample prediction. In practice, models that pose more material model risks are subject to stricter requirements. Additionally, there may be mitigating factors that help decrease the risk, for e.g., the level of manual oversight over the AI/ML outputs. Some risks may be mitigated by having human-in-loop and/or human-over-the-loop in the decision-making process. In cases where the AI/ML output is one of many inputs to a human, the risk is lower and alternatives may be acceptable in situations where there is extensive (external) evidence of the reliability of the algorithm, and adequate internal testing to demonstrate its reliability in the specific context.

B. Post-hoc Explainability

We agree with the premise that the importance of conceptual soundness is well established in industry practice. The first and most important review point starts with the step reviewing key assumptions and limitations; and assessing the applicability of the model to use cases in scope.

There is no single post-hoc explainability technique that works for all use cases, rather there are an array of techniques and solutions that can be applied (and are applied), and their usefulness is very much interlinked to the risk of the specific use case. Explainable AI methods are an emerging research area and new techniques are evolving to improve existing methods.

FIs rely on a combination of development processes and post-hoc evaluation and monitoring tools to mitigate the risks of models or decision-making. These processes as previously discussed allow FIs to better identify when a model is not working as intended.

Each post-hoc technique provides useful information but must be interpreted with critical caveats in mind. Transparency towards customers in terms of explanations of the decision outcomes can be achieved with post-hoc techniques, but it requires additional development efforts to provide explanations that can be easily understood by a human.

There are several reviews that have compiled the voluminous work on interpretable ML,[7] and this area is evolving quickly. Additionally, the *IIF Thematic Series on Machine Learning* highlighted several post-hoc methods used by FIs.[8] A summary of some of the techniques listed in our *Recommendations to Policymakers* paper is included on Appendix B, with the caveat that Appendix B was written in 2019 and is only a reference for some of the explainability methods available and should not be treated as a rule. There are multiple methods for FIs to identify and manage risks related to explainability, such as local linear models, sensitivity analyses, use of model risk's independent credible challenge. We classified the range of different techniques by the scope of interpretability, i.e., whether the technique provides global or local interpretability. Global approaches help understand the entire relationship modeled by the trained response function, which are typically approximations or based on averages. Local approaches promote understanding of small portions of the trained response function, e.g., clusters of input records, and their corresponding predictions, and even single predictions. The wide array of techniques includes global surrogate models, local explanations, or contrastive explanations, etc. [9]

[7] (Guidotti, et al., 2018), (Du, Liu, & Hu, 2019), (Rudin, 2018), (Gilpin, et al., 2019), (Molnar, 2019)

[8] IIF, *Machine Learning Thematic Series*: Explainability in Predictive Modeling, November 2018; and IIF, *Machine Learning Thematic Series: Bias and Ethical Implications in Machine Learning*, May 2019. Can be accessed at: https://www.iif.com/portals/0/Files/private/32370132_machine_learning_explainability_nov_2018.pdf and https://www.iif.com/Portals/0/Files/Thematic_Series_Bias_and_Ethics_in_ML.pdf

[9] We discuss some of these methods in more detail in Appendix B.

Supervisors should clarify expectations around both – internal and external – explainability, understanding that the type of information and depth of explanation will vary for each. The need for explainability should be proportional to the risk of the use case. For instance, the type of explanation needed for customer facing use cases that are higher risk, such as credit underwriting, where firms have an obligation to provide reasons for any adverse decision under Reg. B, is different than the one required internally for ML models that are used for development and model risk management. Existing regulatory guidance does not clarify the degree of explainability needed for a model.

In this vein, non-banks (and especially non-bank consumer lenders) should be subject to similar explainability risk management expectations as banks.

Context Matters

When discussing post-hoc techniques and explainable AI, context matters. Explanations vary depending on the stakeholder, the use case objective, and the risk of the model in question. All these need to be considered within a framework based on model governance tools and processes to test, monitor, and govern ML models.

Firstly, explainability risks are not limited to AI/ML models, even traditional regression models can have hundreds or thousands of correlated variables making them difficult to explain. Additionally, banks may use vendor models where the vendor does not share the inner workings of the model, and overreliance on human intervention has historically been problematic as humans rely on subjective reasoning and judgment.

Secondly, a single ML model can have multiple stakeholders: those implementing a ML application, management responsible for the application, the FI's independent control functions, conduct regulators, prudential regulators, and the consumer.

In addition to stakeholder consideration, FIs consider the trade-offs between interpretability and model accuracy. The level of understanding needed is often linked to the management of the risks associated to the usage of a particular model.

Most FIs utilize several post-hoc techniques; some of the most commonly referred to in the use of ML in credit risk are: feature summary statistic (e.g. feature importance measures), feature summary visualization, model internals (e.g. weights in linear models or the learned tree structure of decision trees), and data points (e.g. to explain a prediction of a data point, find a similar data point by changing some of the features for which the predicted outcome changes in a relevant way).

In practice, model developers conduct analysis on the explainability of AI/ML models using post-hoc methods during the model development process. Such post-hoc explainability analyses may be performed and supported through a number of different techniques such as partial dependence plots (PDPs), variable importance analysis, surrogate interpretable models, and other similar visualizations that describe behavior / distribution of individual features and their corresponding impact on the model output.

Surrogate models are a technique that has been widely used by FIs, i.e., using a simpler model to explain another more complex model to approximate the predictions of the underlying model while retaining interpretability. Many surveyed firms reported using LIME, a type of local surrogate model used to explain single predictions of any ML model, in many cases the explainable model used typically was LASSO.

Model usage and risk are also considerations in the level of explainability needed. For instance, credit decisions that result in a customer being declined a mortgage loan may have different implications than the use of ML for marketing campaigns. Where possible, inherently interpretable models should be prioritized for use cases where explainability is critical. For use cases where more complex but less transparent models are most beneficial, regulators should allow firms to implement appropriate controls/guardrails rather than mandating the use of post-hoc techniques, which may be challenging. These approaches can also be a challenge when working with externally sourced (third-party) models from suppliers who are not yet familiar with traditional model risk management frameworks.

Data Processing and Usage

Question 4: How do financial institutions using AI manage risks related to data quality and data processing? How, if at all, have control processes or automated data quality routines changed to address the data quality needs of AI? How does risk management for alternative data compare to that of traditional data? Are there any barriers or challenges that data quality and data processing pose for developing, adopting, and managing AI? If so, please provide details on those barriers or challenges.

Question 5: Are there specific uses of AI for which alternative data are particularly effective?

Question 6: How do financial institutions manage AI risks relating to overfitting? What barriers or challenges, if any, does overfitting pose for developing, adopting, and managing AI? How do financial institutions develop their AI so that it will adapt to new and potentially different populations (outside of the test and training data)?

A. Data Use, and Alternative Data

Machine learning applications rely on large amounts of data, and often multiple datasets, thus it is crucial to understand what data is being used, if it can and should be used, and do an assessment of the potential risks that could arise from the use of that data.

Alternative data can originate both externally and internally data. Alternative data can be collected from online sources, and in some cases include information that is sometimes correlated with finance data. Alternative data can therefore include unstructured data, such as text fields, voice data, and images, and data coming from a 3rd party provider such as data aggregators.

Internal data sources have become easier to track with the emergence of natural language processing (NLP) techniques and algorithms to derive insights out of data.

Appropriately used, data can facilitate new and improved products and services, increase revenue and mitigate risk. When not appropriately governed, certain data practices can damage the company's reputation and cause a loss of customer and client trust. This goes beyond FIs use of AI/ML and is relevant to data practices across any institution. In this vein, non-bank lenders should be held to the same rigorous standards applied to banks to ensure that all consumers receive effective protections on their data. Consistency in examinations across lenders allows federal regulators to detect misuses of alternative data. Having consistent regulation and robust compliance management provides effective guardrails to ensure diverse forms of alternative data are protected, ultimately increasing access to responsible credit and improving financial crime detection and fraud prevention. Additionally, traditional bureau data excludes millions of

Americans from equitable access to credit. Using alternative data can foster financial inclusion for such groups.

The risks of alternative data can be magnified when lenders rely on vendor models created with alternative data to which lenders have little insights into the type of quality of the data being used by the vendor. Thus, alternative data sourced from vendors should be subject to consistent vendor oversight and data governance expectations.

Alternative data, subject to FCRA, should have similar requirements as traditional credit data pertaining to dispute resolution, ability to dispute, accuracy and transparency standards. Additional barriers with the use of alternative data are inconsistency across data sources, lack of full cycle data, gaps from third-party data in how it is gathered, pre-processed, and whether the data is acceptable from the fair lending lens. Thus, data quality is critically important. Data aggregators, technology platforms, and other alternative data providers should not circumvent these consumer protections. All FIs should follow very structured and controlled processes to ensure data quality of alternative data.

According to our 2020 survey, nearly three-quarters (73%) of our sample have established a firm-wide data governance committee as it relates to ML applications. Enterprise-wide data use governance frameworks have been implemented by various FIs to ensure that data is handled and used properly. This includes assessing the risks of new uses of data; while governing bodies are authorized to approve use of the data, they may impose additional controls to mitigate any identified risks. Data use decisions are guided by a set of firmwide data use principles to ensure that all sensitive information of the firm, its customers, and clients is protected, and that the appropriate use of data creates a positive impact for all stakeholders.

B. Siloed Data

Data quality is not a ML-specific issue, however it is an important one. The issues around data quality also apply to traditional models. In fact, ML models can perform better and help overcome missing variables and data quality challenges better than traditional models. While data quality challenges are omnipresent in banking data, the challenges posed from a ML perspective are different. From their experience, this includes the identification and treatment of ‘special values’ in numerical variables and missing values, as well as transforming data into the right form and engineering the appropriate set of features.

Issues related to IT infrastructure and siloed data continued to pose a challenge to FIs. From 2018 to 2019, we saw 178% increase in the number of firms that singled out siloed data as hindering their ability to leverage ML fully. This was particularly challenging for multinational FIs with legacy IT systems where legacy systems presented a challenge of leveraging data and building analytics on top of that data.

C. Model Validation and Model Monitoring

The sophistication of validation and the choice of techniques employed to assess the robustness of ML models vary depending on several factors, including the use case objective, complexity and/or materiality.

The most common validation method chosen was "in sample/out of sample testing" (91%) followed closely by data quality validation[10] (80%) and "outcome monitoring against a benchmark" (79%).[11] In the U.S., "in sample/out of sample testing" and "outcome monitoring against a benchmark was selected at a 100%. This is also the case for model monitoring, where we see a variety of feedback mechanisms and controls, and safeguards to mitigate the risks of ML models.

All U.S. firms in our sample have feedback mechanisms in place for ensuring outcomes are as expected, and to prevent input data and features from drifting over time. U.S. firms indicated that performance monitoring accounts for feedback mechanisms and controls, and that the purpose of model monitoring, the assessment of the monitoring results and subsequent actions to address any model issues, are required in the model monitoring plan that is designed to capture all nuances of a model, including ML models.

Similarly, FIs use a variety of safeguards to mitigate the risks of ML models, and the choice of safeguard(s) are linked to the individual model in question. For example, areas such as marketing are less complex and impactful than credit risk or fraud detection. In the U.S., the most common safeguard chosen was performance monitoring (88%), followed closely by monitoring model accuracy based on thresholds (75%) and third human in the loop (75%).[12]

There are several metrics to monitor ML model performance, and the choice of metrics is based on factors such as the type of task (i.e., regression versus classification), the business objective, the distribution of the target variable, among other aspects. It could include accuracy, false positive rate, generalizability (is model performance consistent over time). Other metrics around speed, resource usage, might need to have to be met for specific business needs.

Data Quality Validation

In our most recent survey, data quality validation was selected by 80% of participants as one of the model validation techniques used to assess ML model robustness. Data quality validation refers to when one or more techniques are used to ensure potential issues with data (such as class imbalances, missing or erroneous data) are understood and considered in the model development and deployment process. Examples of these include data certification, source-to-source verification or data issues tracking.

FIs check the completeness, accuracy, availability and consistency of the data through gauges for repeatability and reproducibility techniques. In some cases, with newer technology in the pipeline, legacy data stores with monthly feeds are being replaced with daily feeds, allowing for more timely data.

Data sources are expected to follow strong requirements when it comes to data quality, data descriptions and metadata, data sourcing, as appropriate. Some firms have put in place automated, semi-automated and manual solutions that can identify data quality issues in datasets.

[10] "Data quality validation" refers to when one or more techniques are used to ensure potential issues with data (such as class imbalances, missing or erroneous data) are understood and considered in the model development and deployment process.
[11] Outcome monitoring against a benchmark" refers to when decisions or actions associated with the ML system are monitored using one or multiple metrics. Performance is assessed against a certain benchmark value of those metrics.
[12] Other techniques listed on our ML Governance survey were alert systems, backup systems, guardrails, kill systems, and other.

Similarly, in our 2020 survey, regardless of the controls selected, all firms reported engaging in due diligence to mitigate bias risk in ML models. The quality and relevance of data is scrutinized to ensure that models ingest relevant data, as data with a clearly understandable relationship to what the model is trying to predict will help mitigate the risk of bias and have a better representation of the sample. As such, many firms noted that when a model developer designs a model, the quality and relevance of data is scrutinized regardless of whether a ML model or other model is used.

Overfitting

Overfitting is not unique to AI/ML applications, nor is model drift. Rather, overfitting is something that FIs constantly have to manage for. Current model risk management guidelines ensure that models are managed appropriately throughout their lifecycle regardless of the methodology.

In a narrow sense, overfitting may refer to a situation where a simpler model has better performance on a held-out test dataset than a given model. This may happen if the given model is overly complex. In a broader sense, overfitting may refer to a situation where a model corresponds too closely, or exactly, to a training data set, and may therefore fail to generalize its predictive power to other sets of data. These types of models may rely on limited or restricted data sets that do not generalize well in the real world, thus may fail to maintain adequate performance over time.

There are various ways that overfitting can be managed through the progressive phases of model development, model validation and model governance. The risks related to overfitting can be mitigated using three broad approaches:

1. Tools to perform automated stability, robustness testing, and overfitting tests, i.e., to execute the tests planned as part of test strategy
2. Continuous model monitoring and improvement as part of post-production activities
3. Model testing using many stratified hold-out and/or out of time data sets

Additionally, one key aspect to tackling the overfitting issue is focusing on producing quality datasets that can be reused in a systematic way. To achieve that, FIs may invest in data fabrication of synthetic data, i.e., create "real-life" usable data that are multi-dimensional depending on the model requirement, for instance.

Additionally, there are standard techniques of cross-validation of the model on out-of-sample populations. Different regularizations can be applied during model training, and model explainability is another approach. With new data, some firms recommend a monitoring program.

Validation and governance practices are relevant, such as explainability assessments, OOT performance measurement. Overall strong change management and ongoing monitoring, which we discussed earlier in our discussion.

Cybersecurity Risk

Question 7: Have financial institutions identified particular cybersecurity risks or experienced such incidents with respect to AI? If so, what practices are financial institutions using to manage cybersecurity risks related to AI? Please describe any barriers or challenges to the use of AI associated with cybersecurity risks. Are there specific information security or cybersecurity controls that can be applied to AI?

With respect to cybersecurity, regulators have emphasized the necessity for a robust governance and risk management framework. As such, FIs maintain rigorous cybersecurity programs designed to protect firms and their clients, support secure delivery of services, be adjustable to address the risks presented by an evolving threat landscape, and meet regulatory requirements, all while remaining technology agnostic and principle based.

Cybersecurity program typically encompasses the governance, policies, processes, assessments, controls, testing, and training efforts required by industry standards and regulators.[13]

With the accelerating change in technology and an increasingly sophisticated cyber threat landscape, firms leverage their broader risk management frameworks to systematically and consistently identify, control, assess, measure, treat, and govern information and cybersecurity-related risks. The use of AI has the potential to reduce risk and make companies more secure.

These cybersecurity programs, alongside with current regulatory (NIST Cybersecurity Framework, NIST A Taxonomy and Terminology of Adversarial Machine Learning, FFIEC Cybersecurity Assessment Tool) and industry guidance (e.g., Financial Sector Profile, Microsoft/MITRE Adversarial ML Threat Matrix), provide sufficient security measures and guidance to address the risks associated with the introduction and development of AI systems. At this time, we do not feel additional controls, or frameworks are required to address the security concerns associated with the usage of AI and ML. FIs remain mindful of new risks, and continue to systematically and consistently identify, control, assess, measure, treat, and govern information and cybersecurity-related risks.

Risk management of AI and ML models should not be treated differently from other forms of technology and should not have its own set of specific standards. Pursuing additional technology specific controls and guidance could create fragmented risk management practices that introduce operational burden.

Dynamic Updating

Question 8: How do financial institutions manage AI risks relating to dynamic updating? Describe any barriers or challenges that may impede the use of AI that involve dynamic updating. How do financial institutions gain an understanding of whether AI approaches producing different outputs over time based on the same inputs are operating as intended?

We are interpreting dynamic updating to refer to models that are trained online, (i.e., in live use, real time) as opposed to models that are retrained offline (i.e., not in live use) often with guardrails. The former is currently not being used extensively by FIs, there are a few use cases that could benefit from models that are trained online.

Our 2020 survey results indicate that all U.S. FIs have implementation platforms that cater for the need to frequently update/change model parameters (62.5%) or are in the process of establishing them (37.5%). Those with implementation platforms already established indicated that their platforms fully support the AI/ML lifecycle including support for model operations with associated model revisioning and promotion capabilities.

[13] Policies and standards, for example based on the Financial Services Sector Cybersecurity Profile, provide establish the administrative, technical, and physical safeguards for protecting firms' technology environments, facilities, and client information.

As the maturity level increases, AI developers, risk owners, process owners, users, become more comfortable and may consider dynamic updating and put it to use. Currently, most use cases rely on supervised learning with monitoring, and periodic updates to the AI models (with associated testing / validation) rather than dynamic updating.

It is necessary to understand new data patterns, and keep the model updated to operate on these. Depending on the risk of usage and complexity of the update, it is critical to understand the change, test it before the new learning is put to use.

In the U.S., MRM guidelines for ongoing model risk management, include model changes and are effective at managing the risks of dynamic updating models. Dynamic updating models require increased automation of controls typically performed by humans in manually updated models. In practice, these can be in the form of more frequent and/or granular monitoring of model outcomes, where human oversight is engaged if and when dynamic updating models breach allowed parameters.

Some FIs indicated that this was highly situational and implementation specific as some models may be designed to be updated on a more frequent basis depending on the use case. Models may be reviewed on a more frequent basis depending on the complexity and materiality of the model, which could result in the identification of limitations/overlays.

Additionally, model implementation platforms can differ based on different business units. Model developers typically provide a model lifecycle management solution that allows the model owners of the ML model to visualize model results and the ongoing monitoring of performance of key model inputs.

In terms of hurdles to adopting dynamic updating models, some firms face significant software engineering challenges involved in developing and maintaining such systems. Some of the management techniques that firms are using are automated processes, testing, performing explainability checks, and robust validation. In many cases, these types of models have a built-in testing for impact analysis and robustness.

Oversight of Third Parties

Question 10: Please describe any particular challenges or impediments financial institutions face in using AI developed or provided by third parties and a description of how financial institutions manage the associated risks. Please provide detail on any challenges or impediments. How do those challenges or impediments vary by financial institution size and complexity?

Our 2020 survey indicated that most U.S. FIs (78%) have principles in place for the use of externally sourced data and aggregate scores provided by third parties, and the remaining firms are currently defining them.

In the U.S., most firms indicated having Group Wide Data Principles that cover external data. Effective third-party risk management processes can control for any increased risk created by using third-party products and services. FIs comply with existing third-party risk management guidance from regulators to develop and manage third-party relationships. Additionally, third party risk can be managed by relying more on sensitivity analysis and benchmarking per SR11-7.

Fair Lending

Question 11: What techniques are available to facilitate or evaluate the compliance of AI-based credit determination approaches with fair lending laws or mitigate risks of noncompliance? Please explain these techniques and their objectives, limitations of those techniques, and how those techniques relate to fair lending legal requirements.

Question 12: What are the risks that AI can be biased and/or result in discrimination on prohibited bases? Are there effective ways to reduce risk of discrimination, whether during development, validation, revision, and/ or use? What are some of the barriers to or limitations of those methods?

Question 13: To what extent do model risk management principles and practices aid or inhibit evaluations of AI-based credit determination approaches for compliance with fair lending laws?

Question 14: As part of their compliance management systems, financial institutions may conduct fair lending risk assessments by using models designed to evaluate fair lending risks ("fair lending risk assessment models"). What challenges, if any, do financial institutions face when applying internal model risk management principles and practices to the development, validation, or use of fair lending risk assessment models based on AI?

Question 15: The Equal Credit Opportunity Act (ECOA), which is implemented by Regulation B, requires creditors to notify an applicant of the principal reasons for taking adverse action for credit or to provide an applicant a disclosure of the right to request those reasons. What approaches can be used to identify the reasons for taking adverse action on a credit application, when AI is employed? Does Regulation B provide sufficient clarity for the statement of reasons for adverse action when AI is used? If not, please describe in detail any opportunities for clarity.

A. Evaluation and Management

The evaluation of ML models for fair lending compliance is similar to the evaluation of traditional models. For any use related to consumer lending lifecycle where there is adverse action, explainability is a mandatory requirement as required by the ECOA. Hence any lack of explainability here (potentially for vendor models or 4th party) can be challenging. Further the quality of explanation produced by vendors and 4th party may not be easily replicated through verification by financial institutions.

There are regulatory challenges to assessing compliance with fair lending requirements. Thus, regulators should consider providing clarity in areas such as the existence of multitude definitions of fairness, on the tradeoffs between reducing disparities and model performance, and on how FIs should choose among models that have different disparities for different prohibited basis groups.

In addition to the performance of statistical analyses, some firms rely on the performance of qualitative fair lending risk assessments performed by compliance officers with expertise in fair lending, data privacy and marketing, for example. This process identifies potential fair lending risks from model use and assesses or requires requisite mitigating controls such as: the adequacy of model explainability, appropriate adverse action reasons, alternative variables, statistical analysis, and model performance monitoring.

B. Bias

The use of ML by FIs has attracted concerns related to the potential for ML models to perpetuate existing biases against or for a particular group. However, ML can also be an opportunity to make systematic corrections when unfair bias is identified in ways that are not possible through more traditional analytics.

Managing this risk in ML is crucial to the safe development and deployment of ML models. Bias can come from the data, from the model, from the parameters selected, and from the outside environment.

There are numerous steps that can be taken by FIs to tackle issues around unfair bias and ethical implications in ML. As with the issues related to explainability, these need to be part of a framework that is based on governance tools and processes to test, monitor, and govern the safety of ML models.

FIs should have principles-based guidelines in place describing factors to be considered in the deployment of ML. Guidelines should outline relevant questions and risks around ethics and discrimination, and FIs should ensure that these kinds of risks are being considered and adequately addressed. In fact, many FIs have created and are operationalizing their own internal high-level principles around transparency, accountability, ethics, and fairness.

Because unfair bias and ethical issues can stem from many entry points, they need to be addressed within the different stages of the data governance process. The most critical step is that of conceptual soundness, which includes identifying where and how bias is present, reviewing key assumptions and limitations, and assessing model applicability. Also important is the need for traceability, i.e., maintaining records of data characteristics, such as data sources and data cleaning, which can help analysis into the outcomes of ML systems.

Another technique to prevent bias, is unbalanced dataset correction, this should be done when the dataset is highly skewed to a certain customer group. Explainability assessments are also another tool to control for unfair bias in models, this process is iterative and should be done every single time the model is generated and monitored for performance.

For optimal results, other factors need to work together in establishing accountability around data ethics in ML, such as the quality of the data, and the diversity and representativeness of system engineers and data scientists.

As previously mentioned, FIs have several processes and mechanisms in place to ensure fair lending practices, including proactively preventing opportunities for unfair bias. In terms of underwriting, FIs ensure that underwriting decisions are based on the borrowers' credit risk profile and not on other factors. For mortgages, for example, model-driven indicative rates for each loan are determined based on the characteristics of the loan (loan size, credit risk, etc.).

C. Sensitive Attributes

Training data is reflective of human history and previous decision-making; thus, dataset bias can come from data samples not being representative of a group of subpopulations. Bias can also come when the data in the training set correlates with certain protected "sensitive" characteristics that cannot be used explicitly, such as race being correlated with area code. Bias can also become embedded in data through the process of data cleaning and transformation. For instance, feature engineering, which creates tailored attributes based on input variables, can impact ML models. The new "features" created may augment and aggregate certain attributes, while minimizing others.

Finally, bias can also come from the outside environment, for instance the current COVID-19 pandemic initially caused a sudden decrease in the model performance of a number of ML models-.[14] The pandemic created a situation that could not be forecasted based on historical data. However, the COVID-19 pandemic did highlight the importance of continuous monitoring and validation to mitigate the risk.

Existing restrictions on the use of sensitive personal information can make it more difficult for firms to determine if an algorithm discriminates based on a protected characteristic. In fact, systems can produce a disparate impact due to the correlations between the variable within a sensitive/protective class and other closely correlated variables, for example, zip code and race.

Non-discrimination laws and data protection laws demand ethics and fairness, and in many cases prevent people from being discriminated against on the basis of certain protected characteristics.

In the U.S., fair lending statutes that prohibit discrimination in lending predates ML. In the U.S. firms must not use prohibited basis data or proxies for discrimination, and the direct utilization of sensitive attributes in model production or development is viewed by many as unlawful under current U.S. anti-discrimination laws. Additionally, as part of the risk assessment procedures, such attributes or anything acting as a proxy for any protected class are excluded before development commences.

Additional Considerations

Question 16: To the extent not already discussed, please identify any additional uses of AI by financial institutions and any risk management challenges or other factors that may impede adoption and use of AI.

Question 17: To the extent not already discussed, please identify any benefits or risks to financial institutions' customers or prospective customers from the use of AI by those financial institutions. Please provide any suggestions on how to maximize benefits or address any identified risks.

Hiring and retaining talent remains an issue for the sector as FIs compete to build multidisciplinary teams formed by AI/ML experts, and business domain experts. The need to build multidisciplinary and diverse teams impacts all industries but is particularly important for FIs as the sector is heavily regulated. The fact remains that it is difficult to find a senior data scientist with a business background that can leverage data science to obtain business insights and address a business problem.

Given the potential for ML to provide a broad range of benefits, any policy action should not constrain the responsible development and innovative use of this technology, rather its adoption and responsible use by FIs should be encouraged.

AI and ML enables FIs to better utilize the available data, gain rich insights into their clients (and prospective clients) and provide services at lower cost due to scale. The use of externally sourced data to supplement internal data could also help expand the business while increasing financial inclusion.

Additionally, AI and ML can also be used to scale up efficiencies and accuracy in areas like signature verification, document classification, identity validation, auto-routing of unstructured messages, to name a few use cases.

Current MRM principles have sufficient flexibility that should be upheld for AI/ML models used for credit determination. All models should be subject to an appropriate control framework that

[14] Bank of England, *How has COVID affected the performance of machine learning models used by UK banks*, February 2021

ensures that appropriate controls are in place commensurate with the risk of each specific use case, regardless of whether AI or ML techniques are used.

The IIF stands ready to partner with the official sector to facilitate dialogue between the industry and policy makers on issues around the use of AI/ML.

Appendix B: Excerpt from IIF Thematic Series Paper: ML Recommendations to Policymakers

Appendix B was written in 2019 and is only a reference for some of the explainability methods available and should not be treated as a rule.

Tools for Interpretability

It is imperative to highlight that each approach has its own limitations, and its usefulness varies depending on the case study. Our paper on *Explainability in Predictive Modeling* published in November 2018 presents a current catalog of the many different techniques that can be used to gain interpretability of ML models.

1. **Feature Importance**

Feature importance measures the effect that a feature has on the predictions of a model by calculating the increase of the model's prediction error after permuting the feature. Features are considered important if permuting their values increases model error, and unimportant if it keeps the model error remains unchanged. In other words, it estimates the variance of the model prediction due to the exclusion of certain individual features.

Feature importance tells what's important, for final or intermediate outputs, but not how it's important,[15] which we typically consider to be the explanation. In some cases, a human can spot check whether the machine considers important features that the human believes are not.[16] But this becomes impractical as dimensionality rises; presupposes a human has this knowledge; would suffer from human biases like confirmation bias, especially if the list of "important" features is long; and misses latent variables and inter-feature effects.

Global feature importance measures the overall impact of an input feature on the model predictions taking into account nonlinearity. Global feature importance is necessarily averaged and thus of limited use to understand sophisticated ML that behave in ways that are not monotonic, let alone linear, off the average. This problem can be reduced somewhat by segmenting the output into "regions," each with its own set of "regionally" important features, but this is likely useful only for simpler models where the number of distinct regions is small and where the set of important features transition smoothly from one region to the next.[17] If the number of regions is large, or if the set of important features transitions discontinuously from one region to the next, the regional approach reduces to local feature importance, which we describe below.

Local feature importance describes how the combination of learned model rules and individual observations' attributes affect model prediction for that observation. Local feature importance suffers from the same problems as local interpretable proxy models, to which we turn below.

2. **Interpretable proxies**

These techniques attempt to mimic a ML model using a linear regression or sparse tree or rules, further constrained to be interpretable. Surrogate models attempt to highlight the salient features

[15] (Rudin, 2018)
[16] (Du, Liu, & Hu, 2019)
[17] (Ibrahim, Louie, Modarres, & Paisley, 2019)

of more complex models. This is done by constructing a simpler model to approximating the workings of a more complex one.

*Global***:** Sometimes, a global interpretable proxy suffices to approximate ML behavior.[18] Some academics suggest a satisfactory global interpretable proxy while others disagree "identifying globally faithful explanations that are interpretable remains a challenge for complex models."[19] In the case of a decision tree surrogate model, the attributes of a decision tree are used to explain global attributes of a complex model such as important features, interactions, and decision processes. Surrogate model can help visualize, by comparing the "visual" decision making process, the important features and interactions to the human knowledge and expectations.

*Local***:** firms can opt to build local surrogate models, which allows firms to approximate the model predictions on particular sub-sections of the data. LIME are local surrogate models, and a method for fitting local, interpretable models that can explain single predictions of any ML model. In order to remain model-independent, LIME works by modifying the input to the model locally. Rather than trying to understand the entire model at the same time, a specific input instance is modified and the impact on the predictions are monitored.

However, local methods can be demonstrably fragile against some irrelevant model differences,[20] meaning models that are globally and locally similar can produce very different explanations, and, conversely, demonstrably invariant against some relevant model differences where randomizing network weights do not appreciably change the local explanations.

Additionally, local methods must be constrained to make it interpretable[21], e.g., while ML may use word embeddings to analyze language, the human-interpretable proxy must treat the language as bag of words –and those constraints make the problem hard and that could produce an explanation that looks questionable to humans *because* of those constraints.

3. Partial Dependency Plots (PDPs)

Individual Conditional Expectation (ICE) and PDP are tools to increase transparency and accountability of complex models. Given the limitations of each, these are typically used together.

PDPs are a global interpretability method. They show the marginal effect of a feature on the predicted outcome of a previously fit model, showing the impact of one or two variables on the predicted outcome. It marginalizes the ML model output over the distribution of chosen features, so that the remaining function shows the relationship between what we are interested and the predicted outcome. The partial function is calculated by averaging out the effects of all other input features.

PDPs are useful tools to display the relationship between the target and a feature and can aid in describing the nonlinearities of a complex response function. One disadvantage of this technique comes when the features and the PDP are correlated with other model features. PDP's assumption of independence is a challenge, as the features are assumed to be independently distributed from the other model features are averaged.

[18] (Craven & Shavlik, 1996)

[19] (Ribiero, Singh, & Guestrin, 2016)

[20] (Ghorbani, Abid, & Zou, 2018)

[21] (Rudin, 2018)

4. Prediction by Prediction Techniques:

These techniques help answer the driving factors for a particular individual. This is the case of Shapley, and Individual Conditional Expectations (ICEs).

a. SHAP Value Analysis[22]

Firstly, with Shapley value explanations predictions can be explained by assuming that each feature is a player in a game where prediction is the payout. It assigns payouts to players depending on their contribution towards the total payout. Players cooperate in a coalition and obtain a certain gain from that cooperation. The feature value is the numerical value of a feature and instance; the Shapley value is the feature contribution towards the prediction; the value function is the payout function given a certain coalition of players (feature values).

With this technique the difference between the prediction and the average prediction is fairly distributed among the feature values of the instance. Shapley value can deliver a full explanation. It is however time consuming to compute and is used primarily when an approximate solution is not feasible. FIs have also indicated that the method can be computationally expensive, given the millions of possible coalitions of features.

b. Individual Conditional Expectations (ICEs)

ICE plots are the equivalent to a PDP for local expectations, a disaggregated partial dependence plot. They provide a type of nonlinear sensitivity analysis where model predictions for a single observation are measured while a feature of interest is varied over its domain. They can help visualize the dependence of the predicted response on a feature for each instance separately.

The PDP is the average of the lines of an ICE plot, where the values for each line can be computed by leaving all other features unchanged, creating variants by replacing the feature's value with values from a grid and letting the ML model make predictions with these newly created instances. The outcome is a set of points for an instance with a feature value from the grid and the respective predictions.

ICEs can uncover heterogenous relationships which is a challenge with PDPs. However, like every technique it has disadvantages, primarily in that it can only display one feature meaningfully, as two features would require multiple overlaying surfaces. The other issue is that when the feature of interest is correlated with others, not all points in the lines might be valid data points. In practice, FIs use both PDP and ICE in combination.

Further techniques that support understanding machine learning models:

Visualization and exploratory data analysis: Can be useful in providing interpretability of the input data.

Sensitivity Analysis, and investigation on hidden layers: Can be useful in providing interpretability at the model level.

Evaluation possibilities:

- function-based: i.e., how sparse are the features and does it look reasonable?

[22] Lundberg, S. M., & Lee , S.-I. (2017). A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach, CA, USA. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

- cognition-based: i.e., what factor should change to change the outcome and what are the discriminative features?
- application-based: how much did we improve the outcomes compared to traditional approaches and are explanations useful?

REFERENCES

Adebayo, J., Glimer, J., Goodfellow, I., & Kim, B. (2018). *Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values.*

Craven, M., & Shavlik, J. (1996). *Extracting Tree-Structured Representations of Trained Networks.*

Du, M., Liu, N., & Hu, X. (2019). *Techniques for Interpretable Machine Learning.*

Fawcett, T. (1989). *Learning from Plausible Explanations.*

Freitas, A. (2015). *Comprehensive Classification Models - a position paper.*

Ghorbani, A., Abid, A., & Zou, J. (2018). *Interpretation of Neural Networks is Fragile.*

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., & Kagal, L. (2019). *Explaining Explanations: An Overview of Interpretability of Machine Learning.*

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Gianotti, F. (2018). *A Survey of Methods for Explaining Black Box Models.*

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). *Global Explanations of Neural Networks.*

Kim, B. (2017). *Interpretable Machine Learning: The fuss, the concrete and the questions.*

Molnar, C. (2019). *Interpretable Machine Learning.*

Ribiero, M., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?' Explaining the Predictions of Any Classifier.*

Rudin, C. (2018). *Please Stop Explaining Black Box Models for High-Stakes Decisions.*

Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual Explanations without Opening the Black Box: Automated Decision and the GDPR.*

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). *Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations.*