SEPTEMBER 2019

IIF MACHINE LEARNING RECOMMENDATIONS FOR POLICYMAKERS

INSTITUTE OF INTERNATIONAL FINANCE

INSTITUTE *of* INTERNATIONAL FINANCE

Contents

INTRODUCTION

The adoption and implementation of Machine Learning (ML) continues to progress and gather momentum in the financial services sector, as it has in other sectors including agriculture, health, and marketing. Within financial services, prominent areas of application have included credit risk and the detection of money laundering and fraud. The Institute of International Finance (IIF) has been analyzing financial institutions' applications of ML in credit risk through various surveys and research papers.

ML can allow banks to operate more efficiently. For example, better data can enable banks to screen customers and transactions more effectively against sanctions lists. Some banks are using ML algorithms to partially automate financial crime investigations, as seen in our ML-AML Report. Combining external and internal data can help banks understand and monitor the risk posed by each customer holistically.

ML can also contribute to global financial inclusion. Democratization of banking remains a challenge with approximately 1.7 billion adults still unbanked as of 2017.[1] ML can reduce information asymmetry and lead to better informed credit decisions, reducing manual intervention. Scotiabank's Chief Risk Officer Daniel Moore indicated that ML was used to identify which credit card customers are likely to pay late, and they were able to engage these customers proactively to help. This resulted in a reduction of arrears by 10%.[2]

However, the use of ML has created the potential for machines to learn from data that reflects human biases, including unconscious ones, and then exhibit and perhaps even amplify those biases. Humans make biased decisions, whether intentionally or unintentionally, caused by prejudice or by limitations in knowledge or experiences. Bias is unavoidable; in fact, most data collected by human processes will always have some bias in it. The main concern, and the motivation for the increased scrutiny of bias in the world of ML, is that misguided correlations could have powerful implications given the automated nature of ML algorithms, and inherent biases can produce data that amplifies the biases already present in society.

Consensus is lacking on a common definition for ML, but one key feature that distinguishes ML models from more traditional ones is their ability to adapt to data previously unseen by the model.

Background

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.[3] Building on this 1st comprehensive survey report, in July 2019 we published our *Machine Learning in Credit Risk, 2nd Edition*, examining the continuing evolution and progress over the last year and a half.[4]

---

[1] Global Findex Database, can be accessed at: https://globalfindex.worldbank.org/sites/globalfindex/files/2018-04/2017%20Findex%20full%20report_0.pdf

[2] Daniel Moore, "*Machine Learning in Risk Management*" remarks at Risk Minds International Conference, Amsterdam, December 5th, 2018.

[3] IIF, *Machine Learning in Credit Risk*, March 2018. Please note that distribution of the full Detailed Report is limited to the official sector and the participant firms. A short-form Summary Report is available at: https://www.iif.com/publication/regulatory-report/machine-learning-credit-risk

[4] IIF, *Machine Learning in Credit Risk, 2nd Edition Detailed Repor*t, July 2019. 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/3525/Machine-Learning-in-Credit-Risk-2nd-Edition-Summary-Report

Beyond the IIF credit risk surveys, in October 2018, we published the *Machine Learning in Anti-Money Laundering Report*, a similar study to the original ML-CR project. This report surveyed 59 firms, the majority of which were also interviewed for the two ML-CR reports.[5]

Additionally, the IIF published two papers under its Machine Learning Thematic Series, addressing common challenges in the use of ML for credit risk and anti-money laundering (AML). The first paper in this series, *Explainability in Predictive Modeling*, was published in November 2018 and was followed by a second paper, *Bias and Ethical Implications in Machine Learning*, which was published in May 2019.[6]

This Paper

This paper presents a set of recommendations developed with the 87 financial institutions (FIs) that participated in our surveys. We hope these recommendations will aid supervisors and regulators as they face the considerable task of aligning different legal, technical, and policy-related perspectives, while also ensuring that these do not stifle innovation.

Section 1 of this report provides an overview of recent supervisory developments that affect the use of AI in finance.

Section 2 presents a list of industry developed recommendations clustered in three main groups: (i) avoid overregulation, (ii) support adoption, and (iii) ensure a level playing field.

Finally, the appendices cover a range of topics, a framework for adapting bias and ethical implications for ML including examples of current best practices applied to ML, and a discussion on interpretability techniques.

[5] 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

[6] 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

OVERVIEW OF RECENT DEVELOPMENTS

We are supportive of regulatory efforts to promote responsible adoption of ML. FIs share the view that ML driven decisions should be fair and ethical, and the use of ML should not lead to a loss of accountability.

Our 2019 Machine Learning in Credit Risk study finds that the adoption of ML in credit risk modeling and management has increased significantly in the last year, in particular with a sharp increase in the number of FIs running pilot projects. Although the number of FIs using ML in production has only risen modestly, the sophistication of these ML models and the breadth of application across customer segments has increased significantly.

In 2018, applications were found to be primarily for credit decisioning in retail portfolios with some credit monitoring in the large corporate segment, whereas this year we see a sharp increase in usage for small and medium-sized enterprises (SME) portfolios.

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.

However, there have been substantial changes around key challenges over the past 12 months. Supervisory understanding of new processes was the most common challenge in this year's study, with nine times as many survey participants citing it as a challenge compared to 2018. This gap is likely attributable, in part, to the developing awareness by FIs of their regulators' understanding of ML.

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. Developing an appropriate policy environment to support responsible ML requires a more collaborative effort between the industry and policymakers. Therefore, the goal of this paper is to provide recommendations to policymakers to guide them as they consider future policy approaches.

Recent Supervisory Developments

Supervisors are staying abreast of how Artificial Intelligence (AI) is changing risks in financial institutions, and many are examining ways in which AI advances can be used to help them supervise more effectively. For instance, in a recent speech, James Proudman of the UK Prudential Authority discussed the application of advanced analytics in prudential supervision.[7] Similarly, in August 2019, the Bank of England published a staff working paper on ML explainability in finance.[8]

Additionally, many FIs and supervisors have started working on high-level governing principles that take into account fairness, bias, transparency, and accountability. A very well-developed example is the Monetary Authority of Singapore's FEAT (Fairness, Ethics, Accountability, and

[7] https://www.bankofengland.co.uk/-/media/boe/files/speech/2018/cyborg-supervision-speech-by-jamesproudman.pdf

[8] Bracke, P., Datta, A., Jung, C. and Sen, S. (2019). *Staff Working Paper No. 816: Machine learning explainability in finance: an application to default risk analysis*. [online] Bank of England. Available at: https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf?la=en&hash=692E8FD8550DFBF5394A35394C00B1152DAFCC9E

Transparency) Principles.[9] For many FIs the pragmatic approach taken by the Singaporean supervisor is useful in terms of implementation. One key characteristic of FEAT is that governing principles are not intended to replace existing internal governance frameworks, rather the intent is for them to serve as a foundation when making ML-driven decisions.

We see an increase in the development of such principles in other regions as well. For instance, the European Commission's High-Level Expert Group for AI finalized its Ethics Guidelines for Trustworthy AI in April 2019 and published in June 2019 their Policy and Investment recommendations for trustworthy AI.

Likewise, the Privacy Commissioner for Personal Data in Hong Kong has produced a report on Ethical Accountability Framework for Hong Kong, and the Singaporean Personal Data and Protection Commission published in January 2019 their proposed model AI governance framework.

Given the steep learning curve of the technology, it is unsurprising that some supervisors are backlogged and struggling to keep up in the fast-moving space. We believe it is appropriate for regulators to continue refining their knowledge about how FIs are using ML, to monitor risks as they arise, and understand how firms are mitigating unnecessary risks.

Additionally, we support the development of high-level principles that address the issues of fairness, ethics, accountability and transparency. However, in our view this should be a coordinated effort across agencies and jurisdictions. We are already seeing a proliferation of frameworks with slightly different recommendations, but which promote the same underlying values. Instead, it may be beneficial to both FIs and supervisors to promote consistency in approaches.

Industry Response to Recent Developments

Financial institutions were developing these governing principles before regulators, and some institutions have been successfully operationalizing these principles. This requires executive leadership to improve communication between technical, business, and control functions, and to clarify the organization's social and ethical values. Many firms are working with the heads of their various business units to identify key case studies in order to operationalize these principles. For many firms, this has been an opportunity to improve the view of risk resulting in better credit access for customers, and by making processes more efficient (e.g., through automation or by leveraging ML to identify trends in large data sets that humans would not be able to spot).

Given that the Monetary Authority of Singapore (MAS) is currently the only financial supervisor that has developed high-level governing principles to promote fairness, ethics, accountability, and transparency (FEAT) specifically catered to the financial services industry, and that they were the front-runners on such development, many FIs have adopted their principles.

However, as previously stated it would be beneficial to both FIs and supervisors to promote consistency in approaches. For example, this could be achieved at the international level through endorsement and adoption of the G20's May 2019 AI principles which were endorsed by the Organization for Economic Cooperation and Development (OECD) and based on the OECD's own principles published in May 2019.

[9] More information on Monetary Authority of Singapore FEAT can be accessed: http://www.mas.gov.sg/News-andPublications/Media-Releases/2018/MAS-introduces-new-FEAT-Principles-to-promote-responsible-use-of-AI-anddata-analytics.aspx

FIs have taken a cautious approach to the use of ML. Data protection, security, and integrity are already a key part of the design process for banks. Statistical models used for credit decisioning are already subject to model governance, model risk frameworks, and fair lending assessments. FIs can adapt current governance and risk management frameworks to develop approaches to ensuring the ethical use of new technologies such as ML.

POLICY RECOMMENDATIONS

The recommendations presented in this section are *algorithm agnostic*, i.e., they do not focus on specific AI or ML methodology. Rather these are applied to the design, application and use of AI and ML in general.

Our objective is to provide guidance on the key issues and measures that should be addressed by regulators and supervisors.

Theme 1: Avoid Overregulation

Assess current regulatory framework, identify gaps and remove overlaps

Principle 1: Regulatory initiatives should (i) consider how other relevant existing regulations, for example, around privacy and data protection interact, ensuring clarity and consistency, and (ii) assess adequacy of existing requirements before considering new ones.

A wide range of existing regulations apply to the adoption of ML, such as privacy protection regulations, internal governance requirements, and outsourcing rules.

Regulatory dimensions that may affect the use of AI will largely go beyond the remit of financial regulators (and include data protection officers, for example, or be subject to standards that have been designed to apply across sectors). Some degree of coordination and consistency between these various dimensions therefore seems paramount to ensure there is no inefficient duplication (or worse, inconsistencies) that would drown institutions into overwhelming and sterile compliance efforts.

Therefore, before considering the enactment of new (specific) regulations, authorities should assess if the current regulatory framework is adapted for AI adoption, removing any inconsistency or overlap found and providing guidance on how to comply with existing legal requirements in an AI environment.

Once authorities have reviewed the interaction of the existing regulatory framework with AI/ML applications, they should ponder if a clarification of the current regulatory framework according to principles in theme 2 would be enough or, on the contrary, some regulatory intervention is needed, in which case principles 2 and 3 should be taken into consideration.

Implementing controls based on materiality

Principle 2: Regulatory initiatives should take a risk-based approach and guide organizations to determine appropriate controls that are commensurate with the materiality of each specific use case.

Materiality could, where appropriate consider factors, such as:

- the extent to which ML is used in the FI's decision-making,
- the complexity of the ML model used,
- the degree of automation involved,
- the severity and probability of impact on different stakeholders,
- the monetary and financial impact,
- the level of human involvement,
- the impact on regulatory compliance,
- potential cyber security risks, or
- potential reputational impacts

Taking these into account, regulatory initiatives should be aimed at ensuring highly material ML solutions have a positive impact and reducing obstacles to experiment with or apply new ML techniques. It would be FIs' responsibility to calibrate their actions and requirements under their own internal framework.

To illustrate, in the financial context, a single ML model may have multiple stakeholders with their own unique use cases. Thus, the controls should be commensurate with the materiality of each specific use case. For instance, one single ML model may have several different types of stakeholders, for instance: those implementing an ML application, management responsible for the application, the FI's independent control functions, conduct regulators, and prudential regulators. Thus, the impact of an autonomous decision, such as whether a loan gets processed, would be of interest to not only those tasked with implementing the ML application (in order to understand outliers), but also to a conduct regulator. However, other stakeholders may be more interested in how the model works more generally.

Remove obstacles to access data, technology and experiment

Principle 3: Regulatory initiatives should remain dynamic, technology-neutral, and future-proof. Regulators should not impose overly prescriptive requirements, rather they should offer a level of strategic tolerance.

Compared to other industries, early adoption of ML by the financial industry had not been fast-paced, slowly growing from more advanced pilots to deployment in smaller portfolios. However, in the last 18 months we have seen a surge in the use of ML in credit risk and AML, an encouraging development for the industry, which has resulted in developments outside of the retail sector to other portfolio types such as SMEs.

It is of upmost importance for supervisors and regulators to encourage innovation and show tolerance as banks start using new technologies.

Theme 2: Support Adoption

Principle 4: Authorities should support firms in their transition, as FIs update internal governance structures and measures to ensure the robust oversight of the FI's use of ML. This entails establishing a robust governance structure that supports the identification, monitoring, and end-to-end oversight of the FI's use of ML models. Where appropriate, governance should incorporate existing risk management frameworks, control structures, and review processes.

FI's already have existing risk management frameworks and risk control measures that are being assessed and adapted to ensure that new risks and responsibilities related to the use of new technologies are taken into account.

This builds on Principle 1, as FIs should manage the risks of deploying ML by applying governance that is structured to ensure that appropriate controls are in place that are commensurate with the materiality of each specific use case.

Given applications of AI/ML may be developed across different functions of the organization, the governance structure should support identifying, monitoring, and managing the risks of AI/ML through the following three pillars:

- Establishing firm definitions/categories of AI/ML to ensure applications of AI/ML are defined in a consistent manner across the organization
- Maintaining inventory to capture relevant tools, models, etc., to ensure the organization is able to identify and maintain record of all relevant applications of AI/ML
- Ensuring relevant AI/ML tools/models have appropriate end-to-end oversight

ML-based products that perform tasks that fall under regulatory model governance structures, such as SR 11-7[10], should follow those same general modelling best practices guidelines. While some tweaks may be needed on the periphery, regulatory initiatives should not separately replicate or create new model governance standards.

The following recommendations put forward in Principle 4, Principle 5, and Principle 6 deserve particular attention when deploying ML:

- Establish a robust governance structure that supports the use of ML models
- Have controls in place to ensure there are sufficient guardrails to contain ML techniques from going beyond what's an acceptable result
- Maintain internal guidelines to ensure additional risks posed by using ML are mitigated accordingly

In our recent paper on *Bias and Ethical Implications of Machine Learning,* we provided an illustrative example of a simplified governing process for the use of ML. A brief summary is provided in Appendix A.

Principle 5: Given ML-based processes may not always result in behaviors that are understandable to humans or may vary in their level of explainability, regulatory initiatives should require firms to i) have controls in place to ensure there are sufficient guardrails to constrain the ML technique from going beyond acceptable results, and ii) maintain a set of internal guidelines to ensure that any additional risks posed by using ML are mitigated accordingly.

The guardrails/control framework around the process needs to consider how the FI is comfortable that it has contained the risks of a particular process when these less explainable ML techniques are utilized.

- With less explainability, how do you get comfortable that your guardrails/controls are adequate?

Additionally, some material risks may involve aspects of the product beyond the traditional control framework. For example, given that ML techniques generally involve delegating some responsibility for the explicit behavior of the system to the ML algorithm, how does the FI ensure that the actions taken by their product are consistent with the ethics that would be required of a human design?

- With less explainability, how do you make sure you are covering other risks such as:
    - How do you know your process doesn't yield solutions that have ethical/discriminatory bias?
    - How do you feel comfortable that your process is not more prone to manipulation/cyber-attack?
    - How do you feel comfortable that your process cannot engage in manipulative/collusive activity?

[10] U.S. Federal Reserve, *"Guidance on Model Risk Management"*, April 2011

As outlined in Theme 1, AI/ML adoption should not be stifled through overly restrictive regulatory/supervisory approaches. Machine Learning can improve the efficiency and accuracy of financial modeling by, among other factors, using data to systematically perform effective variable selection as well as to capture non-linear relationships between the predictors and target variables. As a result, ML-based processes tend to be complex, making mechanistic, intuitive explanations of the role played by each input variable difficult to state simply.

The IIF has discussed explainability in our *"Explainability" in Predictive Models* paper. Although ML models are typically more complex, technical approaches can aid in identifying which factors are important for making predictions and in inferring how the ML models operate.

Principle 6: In most cases, existing model governance practices are suitable for ML-based processes. However, where appropriate, 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, and FIs should ensure that these kinds of risks are being considered and adequately addressed.

Regulators already give guidance on oversight of model behavior, and in most cases, it is expected that existing model governance practices will be applicable to ML-based processes and regulatory initiatives should not separately replicate or create new model governance standards.

For cases in which it doesn't, FIs should have principles-based guidelines in place that describe factors FIs should consider in their deployment of ML. These principle-based guidelines serve to address the additional questions/risks posed by ML and may include factors, such as:

- How do you know your process doesn't yield solutions that have ethical/discriminatory bias?
- How do you feel comfortable that your process is not more prone to manipulation/cyber-attack?
- How do you feel comfortable that your process cannot engage in manipulative/collusive activity?

There may be many acceptable routes to deploying ML appropriately while providing a flexible approach to innovate. Such principle-based guidelines should be reviewed and/or updated on at least an annual basis to ensure that they remain relevant, dynamic, and useful.

Theme 3: Ensure a Level Playing Field

Sector and country boundaries are blurring, an international and cross-sectorial level playing field is needed.

Principle 7: Regulatory initiatives should avoid regulatory arbitrage, and ensure a level playing field between all players. Guidance should apply to financial *activities* not financial *entities*.

The emergence of bigtechs, fintechs, shadow banks, and other players on the financial services scene does not simply raise competition issues. They also create the risk of regulatory arbitrage. Having a sound data management framework (of procedures, controls, and best practices) has been a cornerstone of banking operations. FIs have had to demonstrate very robust risk management and governance processes to the regulatory and supervisory community. In a dynamic environment with new players dealing with financial data, it is critical that all participants can emulate the sound standards that banks already deliver and achieve the high security standards that customers have come to enjoy.

Therefore, all players providing bank-like services, including the broad range of financial applications that aggregate data, provide valuable insights or initiate transactions, as well as in investments and the wholesale and corporate banking markets, should be held to the same standard. Customers can benefit from greater innovation, competition, and access to financial services. However, accompanying risks must be appropriately identified and addressed to ensure customers are protected at all times and the stability and integrity of the financial system is preserved.

In the area of data and machine learning, big technology platform firms have the advantage of immense scale and the benefits this brings in data gathering and advanced analytics. For example, Amazon invests approximately $22 billion annually in technology research and development while Facebook has 2.4 billion customers. In a marketplace driven by data and new machine learning analysis, this brings a level of competitive asymmetry difficult to grasp. However, when it comes to controls for AI, the financial services industry has strong model risk management frameworks as well as IT risk and controls as a starting point. New entrants do not have the same experience and are not regulated and supervised to the same degree.

Principle 8: Regulatory initiatives should support global and cross-sectorial harmonization of standards, and the specificities of the banking industry should be taken into account in those situations where a one-size-fits all approach is not suitable.

Principle 8 is particularly important given the different initiatives in development around the globe. We encourage regulators to be aware of regulatory initiatives internationally, and we support their alignment where appropriate.

Given different jurisdictional idiosyncrasies we acknowledge that this may not always be possible but achieving a basic level of convergence should be a key objective. For instance, Hong Kong has been moving forward in this area, as well as the United Kingdom with only limited visibility.

APPENDIX A: ADAPTING BIAS AND ETHICAL IMPLICATIONS OF MACHINE LEARNING FRAMEWORKS TO MODELLING

The following data mining governance process framework serves only as an illustrative example; it is entirely appropriate for FI's to have different designs and applications may be different on a case-by-case basis. FIs agree on the importance of ensuring that proper safeguards or technical considerations are taken into account when, for instance, choosing data, or analyzing the assumptions, limitations and weaknesses in the model. This is covered in more detail on our paper *Bias and Ethical Implications in Machine Learning*, which was published in May 2019.

Conceptual Soundness: Critical step of identifying where and how bias and inaccuracies are present, reviewing key assumptions and limitations, and assessing the applicability of the model to models in scope.

Data use: Since ML applications rely on large amounts of data, and often multiple datasets, there should be an understanding of what data is being used, if it can and should be used, and an assessment of the potential risks that could arise from the use of that data. This entails understanding where the data originally came from, how it was collected, how it was moved within the FI, and its accuracy. In instances where the origin of data can be difficult to establish, FIs should assess the risks of using and managing such data. When appropriately governed, data can facilitate new and improved products and services, increase revenue, and mitigate risk.

Data Governance Strategy: Datasets generally require preparation before they can yield useful insights. Therefore, a robust data governance strategy should focus on the data landscape, reference data, and data quality. This entails addressing missing or incomplete values, improper formatting, or other issues that can make processing difficult. When multiple datasets are joined is also important to ensure the integrity of the dataset. Finally, human intervention should be considered, i.e., if a human has applied labels, or edited the data.

Modeling: The modeling process of ML systems is a continuous process of learning. It typically includes choosing and training an ML algorithm, tweaking it, and validating it on holdout data. Algorithms are applied for analysis, results are examined, and algorithms are reiterated until a model that produces the most useful results appears.

Outcome analysis and controls: The results of the algorithm could be subject to clear 1st, 2nd, and 3rd level controls in the same way that other business activities are. The precondition to an effective control framework would be to determine a "benchmark," i.e., to agree which factors and results decide if the algorithm produces biased results.

Tuning and monitoring: Establishing monitoring and reporting systems to ensure that management is aware of the performance and issues related to the use of ML is needed. Proactive documentation of the steps that have been taken to select datasets as well as their source is key. In the same context, audit trails are necessary for the algorithms that are applied, and to show change that have been made to them.

Additionally, many firms remark that validation would need to be monitored, including by model governance groups, over time, to ensure performance degradation is spotted and corrected.

Box 1: EXAMPLES OF MODELING BEST PRACTICES APPLIED TO ML

The following list reflects examples of current best practices that the industry has used over time that are relevant to ML applications.

1. Use different datasets for training, testing, and validation. It is considered best practice to split a large dataset into subsets for these purposes, i.e., train a model using the training data subset, test on model accuracy using testing data, and validate using the validation dataset.
2. Select useful, affordable, legal, correct independent variables.
3. Go far enough back into the past but not too far back (bias/variance trade-off).
4. Explore a rich set of models – nature is not simple and there are often relevant non-linearities that need exploration.
5. Fit each model using reliable software and sensible optimality criteria. R-squared, log-likelihood, and Somers-D are often convenient but very often miss the key idea that really matters – making good decisions.
6. Document so it is reproducible, and accountability can be traced. FIs' decision-making processes should be documented to ensure that data is appropriately stored and retained for the allowed durations, and that data can be used in the future as a training set.
7. Continue to challenge and ensure that what you observe matches what you predicted.
8. Act on the patterns of discrepancies.

APPENDIX B: INTERPRETABILITY TECHNIQUES

Before listing the existing approaches to explainability, it's imperative to note three main points of consideration.

Firstly, if we insist on a global mechanistic understanding – meaning how the ML maps from inputs to outputs – then we're constrained to the simplest of methods: linear regressions or sparse (meaning few features) trees or rules that, to be interpretable, may need to be further constrained, e.g., to be monotonic.[11] Even relatively simple trees are not interpretable.[12]

However, machines can learn things beyond human comprehension, and ML has the potential for increasing FIs' efficiency and effectiveness.

A number of reviews have compiled the voluminous work on interpretable ML.[13] The techniques fall into four categories, that purport to tell us how, for a particular outcome, different features could have produced a different outcome – which is how humans usually explain their actions. [14]

	Global	**Local**
Feature importance		
Interpretable proxy		

The difference between *intrinsic transparency* and *post-hoc interpretability* is relevant. *Intrinsic transparency* constrains to the simplest of methods, while *post-hoc interpretability* extracts information from learned (more complex) models and attempt to highlight the salient features of the model.

While the interpretable ML research is commendable and should continue, each technique provides useful information but must be interpreted with critical caveats in mind. The techniques described below are post-hoc interpretability techniques, i.e., applying interpretability methods after the training.

Secondly, 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.

Thirdly, interpretability needs to be distinguished in terms of three main components:

- Model: a very simple model that a human can fully understand
- Component level (input): each part of the model has an intuitive explanation
- Algorithm: factors that influence the decisions made by algorithm are visible

[11] (Freitas, 2015)

[12] (Kim, 2017)

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

[14] (Molnar, 2019)

Tools for Interpretability

It's 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 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 and also presupposes a human has this knowledge. Additionally, the model may suffer from human biases like confirmation bias, especially if the list of "important" features is long, and a human may miss 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 an ML model using a linear regression or sparse tree or rules, further constrained to be interpretable. Surrogate models attempt to highlight the salient features of more complex models. This is done by constructing a simpler model to approximate the workings of a more complex one.

Global: Sometimes, a globally interpretable proxy suffices to approximate ML behavior.[18] Some academics suggest a satisfactory globally interpretable proxy while others disagree, arguing that

[15] (Rudin, 2018)

[16] (Du, Liu, & Hu, 2019)

[17] (Ibrahim, Louie, Modarres, & Paisley, 2019)

[18] (Craven & Shavlik, 1996)

"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 models can help visualize, by comparing the "visual" decision making process, and the important features and interactions to the human knowledge and expectations.

Local: firms can opt to build local surrogate models, which allow firms to approximate the model predictions on particular sub-sections of the data. Local Interpretable Model-Agnostic Explanations (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 could produce an explanation that looks questionable to humans *because* of those constraints.

3. Partial Dependency Plots (PDPs)

Individual Conditional Expectation (ICE) and PDPs are tools to increase transparency and accountability of complex models. Given the limitations of each, they 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. The method marginalizes the ML model output over the distribution of chosen features, so that the remaining function shows the relationship between what we are interested in 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.

4. Prediction by Prediction Techniques:

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

---

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

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

[21] (Rudin, 2018)

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. The method 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, and 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?
- 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?

[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

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*.

Natalia Bailey
Associate Policy Advisor, Digital Finance
nbailey@iif.com

Other Contributors

Brad Carr
Senior Director, Digital Finance
bcarr@iif.com

Marcus Wimalajeewa
Intern, Digital Finance
mwimalajeewa@iif.com