NOVEMBER 2018

EXPLAINABILITY IN PREDICTIVE MODELING

MACHINE LEARNING THEMATIC SERIES PART I

INSTITUTE OF INTERNATIONAL FINANCE

INSTITUTE *of* INTERNATIONAL FINANCE

Table of Contents

Executive Summary

Against the backdrop of immense progress in Artificial Intelligence (AI) on many fronts, the fear of Machine Learning systems delivering outcomes that are unexplainable is a common concern for banks and their supervisors.

In the IIF's survey on firms' adoption of Machine Learning in credit risk, firms indicated that as they implement these techniques, they are faced with the challenge of tracking and explaining asset-level outcomes. Some indicated that this challenge manifests as a trade-off between interpretability and model complexity, although this trade-off is also present in existing traditional modeling techniques.

Explainability and the technical descriptions of explainable models are diverse and conflicting, suggesting that this refers to more than one concept. In this paper, we make an important differentiation between the activities of interpreting a model, and explaining why a system behaves the way it does. Given the nature of Machine Learning in financial services, the focus should be on being able to interpret results.

In some cases, interpretability is considered synonymous to **understanding how the models work**. In this case understandable models are referred to as transparent, while incomprehensible models are referred to as "black boxes."[1] This understanding raises questions of what it means to be 'transparent', whether we consider transparency at model level, or component level, or at the level of training the algorithm.

In other cases, interpretability is affiliated with **causality** and **transferability**, i.e. explaining "what else the model can tell me." This understanding also raises concerns given that Machine Learning models are typically optimized to make associations, and these associations do not always reflect causal relationships.

With this in mind, in this paper we define **interpretability** as the extent to which a human can understand the choices taken by models in their decision-making process (the how, why and what).

In this context, we focus on three main principles throughout this paper:

- that the type of uncertainty associated with explanation is different than that associated with prediction, and there is an important distinction between explaining and being able to interpret your results;
- that "explainability" as a challenge is not specific to Machine Learning, although there are specific challenges for certain ML techniques, as well as other ML techniques for which "explainability" is not an issue; and
- that Predictive Modeling is an opportunity to make decision making more systematic and accountable.

[1] The term "Black box" is sometimes used to describe closed systems that receive an input and produce an output, but do not offer sufficient insight or explanation as to why or how that output was produced.

This Report

In March 2018, the IIF published our *IIF Machine Learning in Credit Risk Report* (*ML-CR*), where we surveyed a globally diverse sample of 60 firms (58 banks and 2 mortgage insurers) on their experiences, applications, motivations and challenges encountered as they apply Machine Learning (ML) techniques in credit risk.[2]

As the IIF and the 60 survey participants engaged with regulators and other officials to discuss that report's findings, key themes in discussion included the issues of explainability and bias in data. Consequently, the IIF has examined these topics further with the participating firms, producing a three-part series to cover these issues. This is the first paper in this series, focusing on the explainability challenge. The second paper will focus on bias and ethical implications in Machine Learning, and the third paper will elaborate on recommendations to supervisors and regulators.

The first section of this paper discusses the main findings of our *ML-CR Report*, briefly describing the types and uses of Machine Learning identified by surveyed institutions, and highlighting the main challenges uncovered within that Report.

The second section discusses the so-called trade-off between predictive accuracy and interpretability, the important difference between explanatory and predictive modeling, and the difference between explanatory power and predictive power. This section also describes some of the limitations to model interpretation, discusses the importance of the role of the expert, and addresses the necessary challenges in adapting model testing, validation and more broadly model risk management to effectively deal with such techniques. In this context, we also discuss a typical data-mining process and the importance of dividing data into random sets for the purpose of training, testing and validation (i.e. concept of "hold-out" data).

Finally, in the third section, we focus on Machine Learning model interpretation techniques and strategies currently used by experts, and highlight two experiences by different firms of overcoming the explainability challenge.

The IIF has also recently produced a report on *Machine Learning in Anti-Money Laundering*, published in October 2018. Similar to the credit risk study, the IIF surveyed 59 firms (54 banks and 5 insurers), which included a substantial overlap with the same group of firms that were interviewed for the credit risk report. Some of the insights from that Anti-Money Laundering (AML) study are drawn upon for this paper also, but it is not a principal source, noting the very different natures in how Machine Learning techniques are applied across the credit risk and AML fields. Significantly, whereas "explainability" was frequently cited as a hurdle in implementation for use in credit risk, the same was not borne out in our AML study.

[2] 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

1. Introduction to Machine Learning

1.1. What is Machine Learning

The IIF's earlier *Machine Learning in Credit Risk Report* firstly explored and defined some of the key concepts, identifying Machine Learning as part of the wider field of statistics. There is no clear demarcation between Machine Learning and other statistical fields, nor is there a generally accepted definition. For this reason, rather than trying to stipulate a unanimous definition, our earlier report (as well as our *Machine Learning in Anti-Money Laundering Report* of October 2018) identified four main attributes that most ML approaches conform to, as follows:

1. A primary goal of optimizing out-of-sample predictive performance facilitated by well-tuned regularization.
2. A significant degree of automation in the model development process.
3. The use of cross-validation to model relationships in the data, i.e. divide data into random separate sets for purpose of training, testing and validation.
4. Applicable to very large volumes of data (although some techniques also work well on small data sets), in some cases including unstructured data sources.

The first attribute we see very much linked to the need to adapt model testing, validation, and model risk management to effectively deal with these techniques, which we discuss in Section 2 of this paper. The second attribute deals more with the primary goal of Machine Learning, i.e. to get good out of sample predictions, and the use of regularization to penalize excessive complexity, which will form the basis to how we approach the issue of explainability.[3]

1.2. Types and uses in credit risk

Our Credit Risk report's findings are in line with the second attribute, that the primary goal of Machine Learning is optimizing out-of-sample predictive performance, with firms indicating that the most common area of usage was for credit scoring and decisioning, with some application in capital and provisioning (in the form of a 'challenger model' approach) and stress testing.

In terms of its usage for credit scoring and decisioning, three different goals were identified: improving insights from existing data sources in core modeling processes (almost exclusively focused on retail customers); automatic credit decisioning processes (handling applications, such as credit scorecard linked to a decision system); and bringing in new data sources.

Its use as a 'challenger model' has been to serve as a benchmark to the champion model in production, where the ML model might produce new insights about particular correlations in the data, which can potentially be applied to enhance production models.

Machine Learning consists of several different techniques, or as some firms refer to them, *tools*. Two of the most prominent techniques used are **supervised learning** and **unsupervised learning**.

[3] (Varian, 2014)

With **supervised learning** you have a dataset where you have inputs and outputs, and given certain inputs, we want to predict a certain output. In other words, we are trying to predict this outcome variable (output) based on the input data that we have.

With **unsupervised learning**, we typically have a big set of measurements within a large sample. We are looking for correlations within the data, grouping variables and observations, creating clusters, so to say.

Its use in credit modeling was primarily for the functions of (i) model development, particularly for identifying the variables that are the most applicable of the many data items (variable selection) and identifying interactions between variables identified (feature engineering), as well as for (ii) segmentation.

1.3. Main challenges uncovered

To firstly note firms' motivations for using Machine Learning techniques, a 2017 IIF-McKinsey report identified two drivers for the broader digital transformation of the risk management function.[4] The *IIF ML-CR Report* reinforces these two sets of scenarios:

1. a proactive effort by firms to get improved views of risk, with access to better analytics, faster, in order to be a more effective risk management function in and of itself;
2. a push to keep pace with the front-line and evolving customer expectations for immediacy in fulfilment, with a need for speed-to-market in credit decisioning.

In both cases, the expectation is that in order to stay competitive, firms need to deliver faster and better products, supported by a more agile risk function.

Of the main challenges identified in our survey, "explainability" was the most common, although more so for banks with well-progressed pilots than for those who have already implemented these techniques in production models.

Data was the second biggest challenge, itself being both an area of opportunity and challenge. The challenge is that the use of Machine Learning techniques necessitates data to be sufficiently organized and accessible as a prerequisite. Legacy IT systems and data localization requirements can present barriers to developing the depth of structured data for most ML applications. It therefore remains a challenge for firms to create an effective IT environment and data architecture for data consolidation and mining.

Finally, a third challenge identified was "people skills" required to keep up with the evolving techniques and computing power. Several firms reported having a "war for talent," and investing heavily to train staff, and well as competing to attract and retain talent with these skills.

An interesting observation is that similar challenges were encountered in the results of our *Machine Learning in AML Report*, in terms of regulators' expectations around explainability of models, challenges related to data (IT infrastructure), and resources related to "people skills."

[4] (IIF McKinsey, 2017)

2. Interpretability in Machine Learning

In our *ML-CR Report*, firms identified efficiency gains in the forms of better analytics over their data, which combined granular insights into data patterns and relationships with stronger out-of-sample predictive performance. Therein, firms also identify efficiency gains from an improved ability to enhance data, and develop models faster and more accurately.

Machine Learning is not just one technique, but it encompasses entire families of them, from those that allow an algorithm to change the weighting it gives to each data point (boosted decision trees) to those that average together several thousands of randomly generated decision trees (random forests).

There is also no single definition currently for "explainability" and the technical descriptions of explainable models are diverse and conflicting. It's important to differentiate between the activities of interpreting a model and explaining "why" a system behaves the way it does. We argue that given the main purpose of Machine Learning in credit risk, the focus should be on being able to interpret results.

Interpretability can have several meanings depending on the context. In some cases, interpretability might be considered synonymous to **understanding how the models work**. Transparency connotes an understanding of how the model works, and could mean transparency at model level, or component level, or at the level of training the algorithm. This suggests that we are referring to more than one concept.

For instance, transparency at **model** level would mean that a human can fully understand the model, which implies a very simple model (a simple decision tree for instance). But given the limits of human cognition, we believe that neither linear models, rule-based systems, nor Machine Learning models are intrinsically transparent. For instance, a sufficiently high-dimensional model could be considered less transparent than compact neural networks.

Transparency at **component** level – i.e. input, parameter, and calculation – would mean that each part of the model has an intuitive explanation. For example, each node of the decision tree might be linked to a plain text description, and this may infer that models with highly engineered features might disqualify from this notion. However, the weights of a linear model might seem intuitive but may in fact be fragile in terms of feature selection.

Finally, transparency at the **level of training the algorithm** (i.e. algorithmic transparency) means that the factors that influence the decisions made by algorithms would be visible to the people who use, regulate, and are impacted by systems that employ those algorithms.

It is sometimes argued that transparency can enable end-users to game the system: for example, in the case of credit scorecards, applicants could learn that if they have more than two credit cards their score can be affected negatively; so an applicant may simply get a new card after the loan has been approved, improving their score even though the true probability of repayment has remained the same. This is not unique to Machine Learning though, and it may be that an ML model can identify such behavioral responses.

In other cases, interpretability is synonymous with **causality** and **transferability**, i.e. explaining what else the model can tell me. As we discuss below, this understanding also raises concerns given that Machine Learning models are typically optimized to make associations, and these associations do not guarantee to reflect causal relationships.

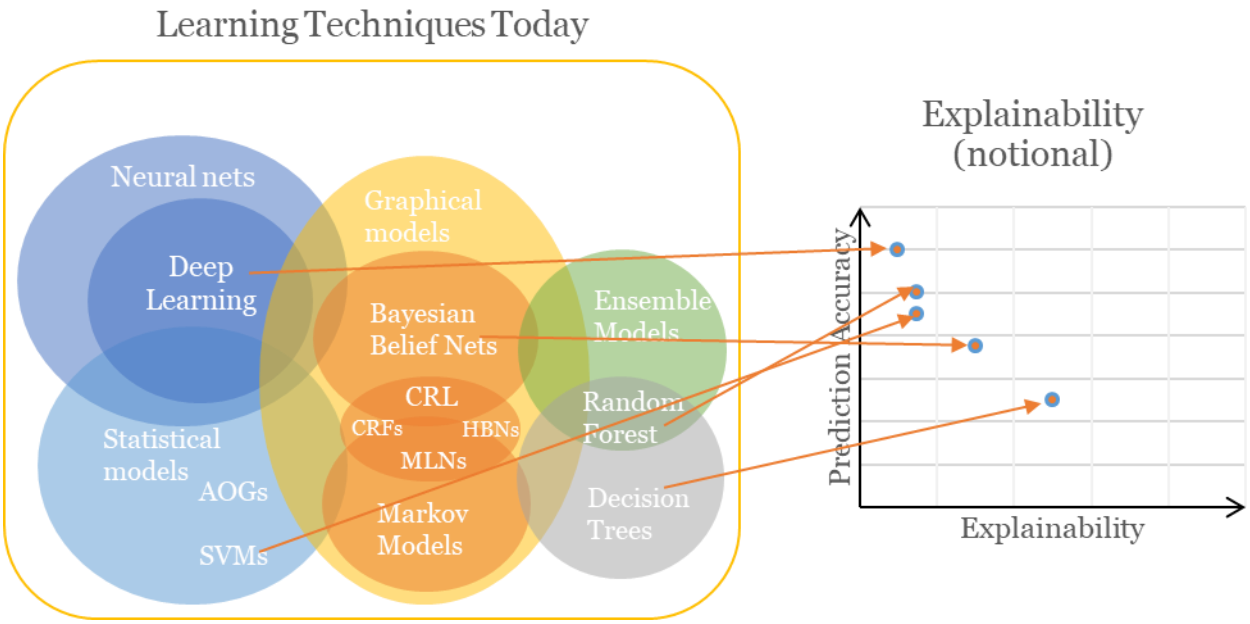
In considering causality and transferability, it is important to distinguish the difference between **explanatory modeling** and **predictive modeling**. Causal explanation and prediction are often conflated, yet each plays a different role. Shmueli (2010) describes Explanatory Modeling as the use of statistical models for testing causal explanations, i.e. testing a causal hypothesis about theoretical constructs,[5] whereas Predictive Modeling is described as the process of applying an empirical model or data mining algorithm to data for the purpose of predicting new or future observations.

Simply put, the focus of explanatory modeling is **causality**, whereas the focus of predictive modeling is *association*. It follows that the primary role of predictive modeling is to generate accurate predictions of new observations.

Secondly, a common misconception is that **predictive power** can be inferred from **explanatory power**, when in fact these are different. **Explanatory power** is the strength of the relationship in the statistical model (i.e. high explanatory power will mean that the relationship in the statistical model is very strong), whereas **predictive power** is the ability to accurately predict new observations. It is possible however to take a "predictive model" and test its "explanatory power," and vice versa.

Machine Learning is not without limitations. A fundamental premise of many of these methods is for humans to understand the parts of the input that are driving the model output. For significant adoption, the role of expert overrides and their judgement will continue to feature prominently. We discuss the integration of ML models with the business process in Section 2.3.

Figure 1: Learning techniques and notion of Explainability[6]



Many experts agree that there is a trade-off between the predictive accuracy of a model and model interpretability. Put simply, a linear regression is typically easier to interpret but does not have sufficient predictive power for cases of nonlinearity. In the other extreme, a powerful

[5] (Shmueli, 2010)
[6] (U.S. Department of Defense, Advanced Research Projects Agency, 2017)

neural net with millions of parameters can give better predictions but is jarring to interpret. In Figure 1, the *U.S. Department of Defense*'s "Explainable AI" illustrates this point, by plotting current learning techniques and the explainability notion, where it uses predictive accuracy and explainability as two separate axes. "Explainability (notional)" is synonymous with the machine's inability to explain its decisions and actions to users, and "prediction accuracy" with maintaining a high level of learning performance.

2.1. Explanatory versus Predictive Modeling

Let's look now at the primary goal of Machine Learning, which is optimizing out-of-sample predictive performance facilitated by well-tuned regularization.

With this in mind, we define **interpretability** as the extent to which a human can understand the choices taken by models in their decision-making process (the how, why and what). This does not entail explaining in detail how a model works, but rather providing useful information for practitioners and end-users. Machine learning model interpretability techniques and strategies are examined in more detail in Section 3. Most methods listed therein allow for post-hoc interpretability, i.e. to select and train a Machine Learning model and apply interpretability methods after the training.

Statistical modeling has traditionally tested causal theories (i.e. explanatory) and uses correlation-based statistical models such as regression or a path model to capture association to decide whether the causal model is refuted or not. Models built for explanatory purposes are different than those built for prediction; they are after different goals. *Explanatory models* are built to test or quantify causal effect for the "average" record in population, whereas *predictive models* are built to predict new "individual" records (i.e. best predict new data).

Figure 2 shows the main differences between Explanatory Modeling (e.g. traditional statistical models) and Predictive Modeling (e.g. Machine Learning, although it may also include several statistical methods). Typically, regressions are the most popular type of statistical models in explanatory work.

Explanatory modeling and predictive modeling are after different goals: for example, a tree is a predictive method, and the predictors can be explained, but it cannot give a causal explanation.

Box 1: KEY DEFINITIONS

Machine Learning Model shall be understood as a model that broadly encompasses the four main attributes described in Section 1.1.

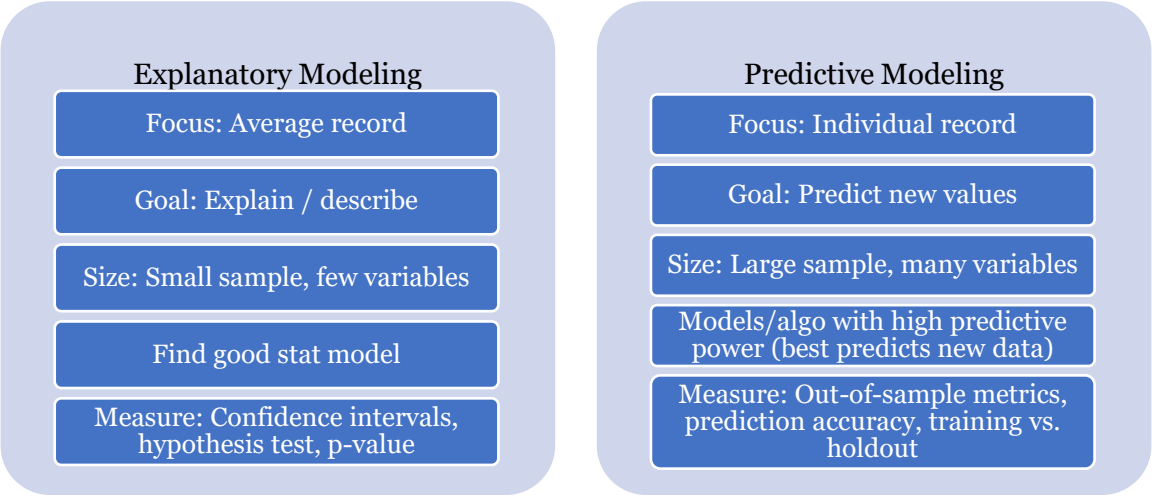
Algorithm is a set of rules that a machine follows to achieve a particular goal.

Dataset: Consists of N tuples of observations, contains the data from which the machine learns. The dataset contains the inputs used.

Prediction: The Machine Learning model "guesses" what the target value should be based on given inputs.

Explanation: To justify the predictions, we use a number of techniques. In some cases, techniques will be *global*, meaning that they would generate an interpretable approximation for the entire model. Other techniques will be more *local* in scope and attempt to rank local contributions for each feature for some observation; this can create reason codes.

Figure 2: Explanatory modeling versus Predictive modeling



As depicted in Figure 2, typically with Explanatory Modeling there is a theoretical model testing a causal hypothesis about theoretical constructs, and so the data fits parameters in a predefined model. Whereas with prediction, that first layer of a theoretical model doesn't exist, instead an algorithm is applied to data for the purpose of predicting new or future observations.

In our view, Predictive Modeling has the potential to make decision making more systematic and accountable. There is a role for both dimensions, and models (whether using ML algorithms or not) will possess some level of each.

2.1.1. Predictive accuracy and Explanatory Power

Earlier we made the point that "explanatory power" cannot be inferred from "predictive power", as they are measuring different things. Explanatory Models use measures such as p-values, R2, goodness-to-fit, and the biggest dangers are typically type I, II errors. Predictive Models typically measure out-of-sample metrics, prediction accuracy, training vs holdout, and the biggest danger is overfitting/underfitting.

Considering both predictive power and explanatory power as two different axes, a type of a bi-dimensional approach, implies changes to model risk management, but also visualizes the trade-off and helps a model developer find the sweet spot in the explanatory/complexity dimension. Depending on the use case, explanatory modeling may not always be needed, and in others predictive modeling can be a short-term solution.

The role and distinction of predictive modeling can help uncover potential new causal mechanisms and lead to the generation of new hypotheses, given that large and rich datasets typically contain complex relationships that are hard to hypothesize. In doing so, it can suggest improvements to existing explanatory models.

What the literature and experience indicate is that the biggest challenge to the use of ML algorithms is not its complexity, but rather the need to adapt model testing, validation and more broadly model risk management to effectively deal with such techniques. Several Machine Learning models can be as describable as a regression model through the partial dependency plots and the individual conditional expectation plots.

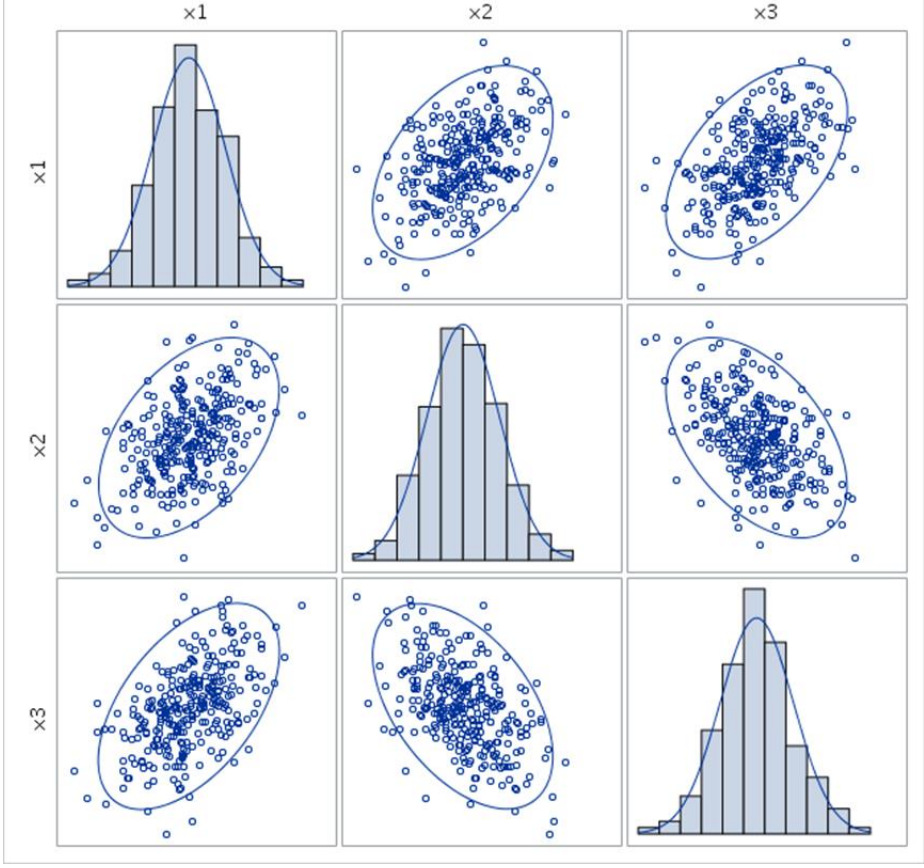
Our first identified attribute (as described in our *ML-CR Report*) was the use of cross-validation to model relationships in data; that is to divide the data into separate sets for the purpose of training, testing and validation. Machine Learning does a great job at predicting within a training set, but what is needed is a model that extrapolates well into the future. This again is linked to the issue of model testing and validation.

2.2. Explainability is not unique to Machine Learning

The issue of explainability is not specific to Machine Learning, or (even) neural nets, but can be present in linear or logistic regressions, as well as decision trees. These can become very difficult to interpret for high dimensional inputs.[7]

For example, even a simple three variable linear additive model can be very difficult to describe intuitively. Consider having 60-day delinquencies, 90-day delinquencies and FICO score in the model. The coefficient for the 90-day variable means the impact of having one more 90-day event, but no additional 60-day events, and no change in FICO score. To illustrate this point further, Figure 3 presents a scatter plot matrix - can any normal person comprehend what the three-dimensional distribution of these data points are? If we were to add a fourth space it would become even more non-intuitive.

Figure 3: Sample of trivariate distributions



[7] (Lipton, 2017)

Linear models are not necessarily more interpretable than Machine Learning models, because if linear models are given high dimensional or heavily engineered features, these models lose transparency at the model level, and at the level of individual components.

Finally, taking this issue more broadly, explainability as an issue is present in any instance in which there is a role for a human expert. Human decision making generally involves some subjectivity, as evidenced in university admission processes in the United States. No matter how you look at it, the university admission process cannot be fully objective, and is linked to several factors such as reviewer criteria (e.g. leadership means different things for person A and person B), and inherent cognitive biases. Therefore, one additional step that should be taken is not to rush to emulate human intelligence. We will explore the topic of bias and ethical implications in Machine Learning on Part II of this thematic series.

2.3. Integrating ML models with business processes

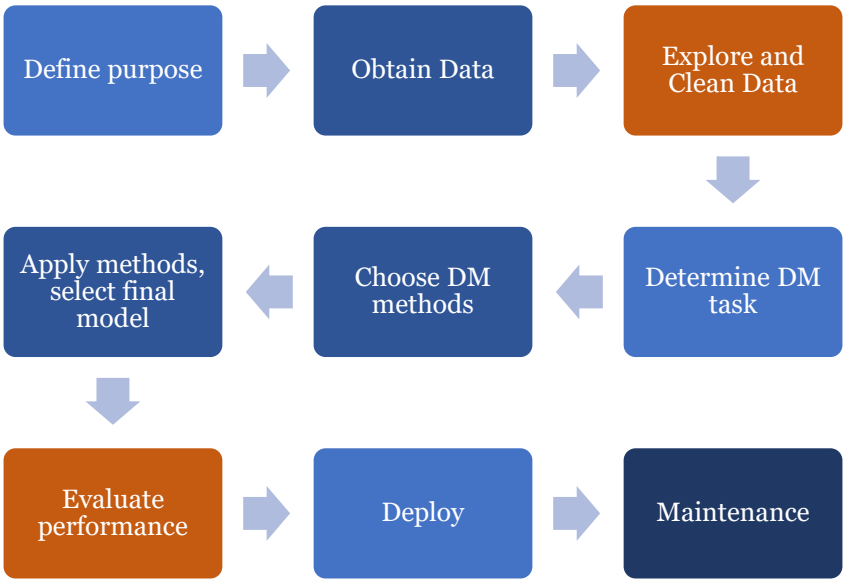
Discussing explainability and Machine Learning goes hand-in-hand with discussing a typical data mining process and its integration with the business process. It is our view that both should be closely integrated.

A typical data mining process is presented in Figure 4, and many of our participating firms using Machine Learning techniques have a similar process in place. It is a large process, that cannot happen in isolated IT departments. In fact, some practitioners argue that several stages of this process that may be more time consuming and detailed oriented than the regular business process. Where data (and IT infrastructure) is an identified challenge, these techniques necessitate data being sufficiently organized and accessible, which in the data mining process involves both obtaining the data, and exploring and cleaning it.

This relates back to one of the attributes of Machine Learning models - the use of cross validation to model relationships in data. Experts and firms that have successfully deployed Machine learning models agree that the (best) practice is to divide data into separate sets for purpose of (i) training, (ii) testing and (iii) validation. Training data is used to estimate a model, the validation data to choose model, and the testing data to evaluate how the model performs (in some cases validation and testing are combined, although experts agree that having three hold-out data-sets is best).

This process of data partitioning will happen early on, during the data exploration and preparation stage. It's very important to remark that, when training the model, these random data-sets are kept separate, and the validation and testing data sets are not used or seen. The idea is to evaluate on a different random hold-out than what was used to train the model. Once deployed, the model will use new data (future data). Poor cross-validation allows "bad" models to make the cut by over-estimating the true lift to be expected in future data.

Figure 4: Typical data mining process



Where Machine Learning techniques form part of a firm's toolkit of different data mining techniques, they do not always know which technique will work for each problem, and therefore they try different sets of techniques. It is an iterative process, where they tweak and evaluate, and once they are done, they choose a final model. That final model is then evaluated for its predictive performance, and that happens by deploying it to a set of data that the model hasn't seen yet (i.e. hold-out data). When it performs according to a predetermined set of requirements, it is deployed.

It is often asked whether using Machine Learning techniques will ensure that we still get answers to questions such as: "Is the model still performing?" This point goes back a bit to the point of "drivers" of risk being stable over time. Predictive analytics learns from the past to predict the future and relies on correlations and associations between inputs and the predictive output, not on causal relationships. It is also true that although correlations are more sensitive in the long term, this does not mean that relying on correlations is negative as they can be a useful proxy for looking at the similarity of things. However, it is important to remember that predictive analytics do not tell us "why."

Machine Learning is typically a predictive tool that builds on correlations/associations, however by examining when it predicts accurately and studying the reasons for the accurate or inaccurate predictions, Machine Learning can give insights on possible explanations, linking results to existing theories and giving ideas for new ones. The use of Machine Learning techniques for predictive purposes does not preclude using linear models for interpretation. However, it's important to remark that the associations learned by Machine Learning algorithms do not guarantee causality. In fact, there are very few causal statistical models, almost all of them are correlative, and it is unusual to have a model that is good at explanation and prediction.

Using these associations learned by learning algorithms to generate hypotheses that could then be tested with an explanatory model is not part of this process, rather it would be an additional step taken. It is important to identify the type of problem that needs to be solved early on, e.g. is the goal predictive or explanatory in nature, as the analysis process will take different avenues

based on the goal. This includes what performance measures are devised, and what types of methods are employed, etc.

For predictive modeling for instance, we should focus on understanding the "correlates of risk" rather than "drivers of risk". The word "drivers" infers a causal question (and is heavily used in credit risk), but there are in fact very little causal statistical models. For example, the assumption that using a consumer behavior model will be stable over time is unlikely, even for a model using logistic regression. The expectation with predictive models (including regression) is that they need to be rebuilt as frequently as needed, and in certain cases for risk models from our participating banks, it is done yearly. We see this as the "maintenance" step in the process.

In terms of evaluating how a Machine Learning model is performing, most firms agree that it is more valuable to calculate the difference between the predictive and observed (i.e. residual), and examine the patterns in the residuals, than use traditional metrics.

Explainability is therefore hardly ever the goal, rather the goal is interpretability. Interpretability can have different meanings and understandings, and in this paper, we define **interpretability** as the extent to which a human can understand the choices taken by models in their decision-making process (the how, why and what). In terms of being able to explain where the uplift comes from over traditional models, we can also do so, for example, by taking the difference of two model scores and putting it into a CART tree and looking at the nodes, with certain nodes predicting higher risk or lower risk.

Finally, it's important to go back to the motivations (discussed in Section 1.3) banks have for using these techniques. This may come from the need for firms to be armed with better analytics in cases of fraud modeling for instance, but is equally important in cases where firms are competing with other firms. Firms need to obtain high performance regardless of whom they are competing against. Although it is true that Machine Learning techniques present their own challenges, there is also an underlying assumption that upper management fully understands how logistic regressions and other traditional modeling approaches work, when this is not always the case.

In our view, we need a combination of causal modeling and predictive modeling in order to make use of data for short-term and long-term actions and planning. It's a cycle. Predictive tools can help with discovering correlations, that can be used to figure out why, and then help improve long-term predictions.

3. Conquering Interpretability

There is rapidly growing literature on techniques to "open up the black box," and to interpret more complex algorithms. Firms are trying to incorporate a number of contemporary approaches to increase the transparency and accountability of complex models.

In this context, the difference between **intrinsic transparency** and **post-hoc interpretability** is relevant. Selecting and training a Machine Learning model that is considered **intrinsically transparent** would imply a simple model, such as a short decision tree. Whereas, **post-hoc interpretability** allows to extract information from learned models, providing useful information for regulators and end-users, i.e. applying interpretability methods after the training.

Some common approaches to **post-hoc interpretations** are visualizations of learned models, and natural language explanations. We may see a variety of different outcomes, such as 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), 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).

Additionally, we have classified the range of different techniques by the scope of interpretability, i.e. whether the technique provides global or local interpretability.

Global interpretability can provide global transparency in Machine Learning algorithms, their results, and the relationship between inputs and target feature. Global transparency helps humans understand the entire relationship modeled by the trained response function, however global interpretations are typically approximations or based on averages.

Local interpretability promotes understanding of small portions of the trained response function, e.g. clusters of input records, and their corresponding predictions, and even single predictions. Typically, local explanations can be more accurate than global explanations due to the likelihood of these being more linear.

3.1. Model Interpretability Strategies and Techniques

Our *ML-CR Report* uncovered that the main use of Machine Learning techniques in credit modeling was for the function of model development, particularly for variable selection and feature engineering. Several firms reported that the use of Machine Learning could make complex features more transparent, which we explore in the FICO example.

As Machine Learning is often used for feature engineering (creation of tailored attributes to be used in strategies or models) there is a need to be able to describe the meaning of the feature and explain their value compared to more traditional attributes/factors. This could be done graphically, through residual analysis or localized partial dependency plots.

Another important aspect and strategy to take into account is how to compare the impact of attributes between models (this is also mentioned briefly in the FICO example). As the use of Machine Learning models expands into more areas, developers will be pressed to produce comparisons with existing models. For simple models, a common comparison is the weight of certain attributes in the model and their contribution to the risk assessment. As the change in loss odds for different values of an attribute or feature of an Machine Learning model can be

opaque, other tools (partial dependency plots, information value, etc.) should be used to make this comparison.

The techniques listed here have proven to increase transparency and accountability of complex models in some manner, and are mostly **model agnostic interpretability methods**, which means that they separate the explanations from the Machine Learning model. One advantage of model-agnostic interpretability methods over model-specific ones is their flexibility.

Another strategy used by a few firms but not discussed in detail in this piece, is the use of automated Machine Learning, i.e. the automation of model fitting. This strategy can include the automation of feature selection, optimization of hyperparameters, comparison of different models, and "ensembling" models. Such solutions already exist and are available via vendors; its use is dependent on the use case scenario. In practice, a business user would ideally work together with a Machine Learning expert to train a Machine Learning model.

3.1.1. Global Interpretability

We first discuss techniques that achieve global interpretability, focus on decision tree surrogate models, Partial Dependency Plots (PDP), and Random forest feature importance. In each case, we discuss the technique, weakness, and how it improves transparency and accountability.

Decision tree surrogate models

We start out with a simple concept, that is surrogate models. Surrogate models are data mining techniques in which a simpler model is used to explain another usually more complex model. Surrogate models are also used in engineering when an outcome is interesting but expensive, and difficult to measure.

The purpose is to approximate the predictions of the underlying model as closely as possible while retaining interpretability. Given a learned function and a set of predictions, surrogate models can be trained to preserve interpretability, often restricted to linear models or decision trees.

In our *ML-CR Report*, we reported that several banks had a "challenger model", in which a Machine Learning model was used to produce new insights about particular correlations in the data, which can then potentially be applied to enhance production models. We saw that challenger models served as a benchmark to the "champion" model in production, firms used it as they did not have to comply with all the requirements for the "champion" regulatory model, and thus the Machine Learning model could be more complex and opaque, allowing the use of non-parametric and other approaches that can provide highly granular insight into data relations. Several banks already indicated in our *ML-CR Report* that they began using such challenger models as part of the validation process for their capital and provisioning models.

The attributes of a decision tree are used to explain global attributes of a complex model such as important features, interactions, and decision processes.

In the case of a decision tree surrogate model, g represents the feature transformations and model, and the surrogate model is a decision tree (htree). The decision tree surrogate model (htree) is used to increase transparency of g by using an approximate flow chart of the decision-making process of g. It also shows important features, and important interactions in g. As such, the decision tree surrogate model can help visualize, debug g by comparing the "visual" decision making process, the important features and interactions to the human knowledge and expectations.

We also discuss K-LIME (K Local Interpretable Model-Agnostic Explanations) plot under Section 3.1.3., which provides a degree of local interpretability to this method.

There are three main advantages:

1. Flexibility: decision tree surrogate models can create explanations for models of nearly any complexity.
2. Intuition: relatively easy to implement and explain to those unfamiliar with ML.
3. Can be easily measured with R squared, in terms of how good the surrogate model is approximating the ML predictions.

Two main concurrent disadvantages are:

1. The surrogate model never sees the real outcome, so it is crucial not to draw conclusions about the data.
2. It is unclear what the best cut-off for R squared is to be confident that the model is close enough.

Partial Dependency Plots (PDP)

Many participants in our ML-CR survey indicated that they use Individual Conditional Expectation (ICE) and Partial Dependency Plots (PDP) to increase transparency and accountability of complex models.

PDP shows the marginal effect of a feature on the predicted outcome of a previously fit model[8], showing the partial dependence as a function of specific values of a feature subset. The prediction function is fixed at a few values of chosen features. It marginalizes the Machine Learning 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.

Partial dependence plots are model agnostic, and a global interpretability measure that can be used to explain response functions of almost any complexity. It can show if the relationship between the target and a feature is linear, or more complex.

Advantages of PDPs include:

- Understanding and transparency are increased by describing the nonlinear behavior of complex response functions.
- Comparisons are enabled of described nonlinear behavior to human knowledge and expectations.
- Intuition: easy to understand and implement.
- Interpretation is clear if the feature and the PDP are uncorrelated with the other model features, as the PDP shows how on average the prediction changes in the dataset.

Disadvantages include:

- A danger 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.
- Limitations on the number of features that can be looked at jointly (two or maybe three), given humans inability to imagine 3-dimensions (discussed in Section 2.2).

[8] (Friedman, 2001)

- Heterogenous effects may be hidden, as it shows the average over the observations. One solution is to look at Individual Conditional Expectation (ICE) curves instead of the aggregated line in order to find these hidden effects.

Random Forest Feature importance

Feature importance was also cited in several instances by firms in our ML-CR survey. This is an example of model specific explanatory technique, rather than a model agnostic technique.

Feature importance measures the effect that a feature has on the predictions of a model. Global feature importance measures the overall impact of an input feature on the model predictions taking into account nonlinearity. Local feature importance describes how the combination of learned model rules and individual observations' attributes affect model prediction for that observation, we discuss the latter one in Section 3.1.2.

Feature importance is measured by calculating the increase of the model's prediction error after permuting the feature. Features are considered important if permuting its values increases model error, and unimportant if it keeps the model error unchanged.[9]

In the case of a random forest, a random forest surrogate model is trained on the predictions of the model, formed by B decision trees. For each split in each tree, the improvement in the split-criterion is the importance measure. The measure is accumulated over all trees separately for each feature. This aggregated feature importance values are scaled between 0 and 1, so that the most important feature has a value of 1.

Advantages:

- Can be used to explain tree-based response functions of almost any complexity
- Conforms to human knowledge and expectations in terms of understanding, providing global insight into a model's behavior.
- Considers interactions with other features; by permuting the feature, it takes into account the feature's main effect and the interaction's effects on the model performance.

Disadvantages:

- Feature importance measure is tied to model error, which is a disadvantage if you want to know how much the model output varies for one feature (ignoring what it means for the performance).
- Needs access to the actual outcome target; permutation feature importance cannot be computed without the actual target.
- If features are correlated, the permutation feature importance can be biased, this is the same as the problem with PDP.

3.1.2. Local Interpretability

The techniques described above provide global interpretability. Below we discuss the techniques that provide local interpretability, promoting understanding of small portions of the trained response function. Typically, local explanations can be more accurate than global explanations due to the likelihood of these being more linear.

[9] (Breiman, 2001)

Individual Conditional Expectation (ICE) plots

Individual Conditional Expectation (ICE) plots are an adaptation of PDP. They create a more localized explanation for a single observation of data using the same ideas as with a PDP. They are 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.

Where PDP are global methods as they focus on an overall average, ICE plots are the equivalent to a PDP for local expectations, a disaggregated partial dependence plot. They 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.

In terms of advantages, ICE shares similar advantages to those listed under PDPs, but they can be even more intuitive than PDPs. One main difference and advantage is that they can uncover heterogenous relationships which were a challenge with PDPs. The identified disadvantages are:

- Can only display one feature meaningfully, given that two features would require multiple overlaying surfaces
- Correlation with other features: when the feature of interest is correlated with others, not all points in the lines might be valid data points

With ICE plots it is difficult to see the average. One suggestion is for firms to use both PDP and ICE in combination.

Leave-one-variate-out (LOCO) local feature importance

LOCO is a variation of the Random Forest Feature Importance measure that we discussed in Section 3.1.1. It allows for calculating feature importance values for any model on a per observation basis. It does so by subtracting the model's prediction for an observation from the model's prediction for an observation of data without an input feature.

It is model-agnostic and allows for local interpretability and can be calculated in different ways. One way is to use a model-specific technique by using a random forest surrogate model, and scaling between 0 and 1 giving the most important feature for an observation of data the value of 1.

In terms of benefits, it allows for local interpretability, and can be used to explain tree-based response functions of varying complexities. It also creates explanations for each model prediction. The disadvantages are similar to those listed for Random Forest Feature Importance.

SHAP Value Analysis[10]

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 Shapley value, is a method for assigning payouts to players depending on their contribution towards the total payout. Players cooperate in a coalition and obtain a certain gain from that cooperation.

[10] (Lundberg & Lee , 2017)

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).

Advantages:

- The difference between the prediction and the average prediction is fairly distributed among the features values of the instance. Shapley value can deliver a full explanation.
- It allows for contrastive explanations by allowing comparing a subset, or a single data point. This is something that other local models don't have.

Disadvantages:

- Time consuming to compute. This is a method that perhaps should be used in cases where an approximate solution is not feasible. It can also be computationally expensive, given that there are typically millions of possible coalitions of features, and the absence of a feature needs to be simulated by random samples. Decreasing the number of samples reduces computational time but increases the variance of the estimate.
- It doesn't work that well for situations in which explanations that involve only a few features are needed.

A recent addition to this set of methods, it attempts to unify these prior attempts at interpreting model output. SHapley Additive exPlanations (SHAP) aims at addressing the issue of interpretability of complex models.

SHAP assigns each feature an importance value for a particular prediction. Its components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties.

The new class unifies six existing methods. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

3.1.3. Global and Local Interpretability

Some methods can give both global and local interpretability. We discuss Decision Tree Surrogate models in Section 3.1.1., which are model agnostic and give global interpretability. Below, we look at K-LIME, a type of surrogate model that gives both global and local interpretability.

K Local Interpretable Model-Agnostic Explanations (K-LIME)

K-LIME is a variant of the Local interpretable model-agnostic explanations (LIME) technique[11], in which local generalized linear model (GLM) surrogates are used to explain predictions of complex response functions.

LIME are local surrogate models, and a method for fitting local, interpretable models that can explain single predictions of any Machine Learning model. LIME tries to see what happens with model predictions if you feed variations of your data into the ML model. It then generates a new dataset consisting of perturbed samples and the associated ML model predictions. This dataset LIME trains an explainable model weighted by the proximity of the sampled instances to the one

[11] (Ribeiro , Singh, & Guestrin, 2016)

of interest. The learned model is usually a good approximation of the ML model locally, but it may not be globally. The explainable model can be a decision tree, or LASSO for instance.

K-LIME defines local regions as K clusters or defined segments rather than simulated perturbed observation samples. For example, local regions can be segmented with K-means clustering, dividing input training data into K disjoint sets. For each cluster, a local GLM model is trained, and K is then chosen so that predictions from all the local GLM models maximize R squared.

In addition, K-LIME can also train one global surrogate GLM (hglobal) on the entire input training dataset and global model predictions. The result is that: intercepts, coefficients, R squared, accuracy and predictions from all surrogate K-LIME models can be used to increase transparency.

The advantages of K-LIME are that they can provide explanations for complex ML models:

- The coefficient parameters of the global GLM surrogate give information about global, average trends. Also, the coefficient of in-segment GLM surrogates can be used to provide average trends in a local region and understand the average direction of the effect of an input feature.
- Reason codes values can be generated with K-LIME for specific in-segment observations, allowing users to understand the approximate magnitude and direction of an input feature's local contribution.

The disadvantages are that:

- Accuracy can decrease when the model is too nonlinear, meaning that wide input data or strong correlation between input features can decrease the quality of local explanations.

This method increases transparency by revealing important input features and their linear trends. It also enhances accountability by creating explanations for each observation in a dataset. It shows the important features and their linear trends around specific records that conform to human domain knowledge and reasonable expectations.

3.2. Overcoming the explainability issue

3.2.1. Case Study: Scotiabank R&D Success Story

One R&D success story is that of Scotiabank. Scotiabank retail modeling teams were asked to optimize new credit acquisitions for retail loans while staying within the prescribed risk appetite. They solved the business problem by creating a customized ML algorithm, which in conjunction with new sources of data, led to better risk discrimination power while enhancing model stability thereby increasing acquisition volumes within the Bank's risk appetite.

Consumer risk modeling typically involves ranking customers by credit worthiness (i.e. how likely/unlikely they are to repay a loan). This is typically done by identifying customer characteristics that indicate risk of delinquency, and then calculating a relative risk score for each customer. Historically, the output of the model has been presented in a type of scorecard. Credit scorecards are a way to present customer risk models in a simple, readable fashion. Further, they are interpretable and implementable in legacy adjudication decision systems.

A scorecard identifies certain characteristics that indicate risk using statistical techniques. Each characteristic is subdivided into a small number of bins and each bin is assigned a number of

score points which is proportional to the risk of that bin. An applicant falls into just one bin per characteristic, and the applicant gets one score for each characteristic. A final score of the applicant is the sum of the points assigned by each bin. Typically, this is done via some type of logistic regression.

Scotiabank has been working on using advanced Machine Learning techniques to predict risk. In this particular case study, Scotiabank analyzed several advanced Machine Learning techniques, including ensemble techniques (e.g. boosting, bagging) and neural networks (including deep learning). Based on their analysis, they focused on an ensemble method called boosting to predict the risk of a customer. Boosting is a Machine Learning technique that has a lot of similarities with stepwise weight of evidence logistic regression (SWOELR), which is a popular methodology for credit scorecards. It is a well-established method with empirical and theoretical support spanning 15 years. Scotiabank made significant modifications to the boosting algorithm to customize it for the intended purpose of credit modeling. The modifications allow the user to tune the complexity of the model and help the user "explain" the model with a scorecard compatible to SWOELR.

Scotia has found a significant lift (up to 70% in some sub-segments) for multiple retail portfolios, which has resulted in a significant increase in Banks' risk adjusted margins while staying within the Bank's risk appetite.

3.2.2. FICO

Machine Learning and new analytical techniques are currently used at FICO[12] for several purposes, with an emphasis on tackling "interpretability" right from the raw data. In this example, we look beyond the model and to the whole process, with the FICO team noting that if the path to get to the model is made transparent and interpretable, allowing for human intelligence to be interjected along the way, the outcome is easier to explain.

FICO is currently working on using ML techniques to streamline, in a methodical and efficient manner, the process of data wrangling,[13] enrichment (including social network analysis, etc.) while creating accurate and explainable results. This involves working to develop tools and methodology that identify and prioritize the "signals" from the data and produce visuals that the data scientist can use to facilitate communication with the business experts so that both develop a comprehensive understanding of the data.

FICO has been investing in developing these capabilities by working with clients to find value in the abundance of data that they already have, augmented by data from other sources.

To start with the end result (typically a model), FICO has created tools/methodology in which data scientists can better understand their ML models – the Machine Learning algorithm does its best to come up with a model, and then the user can interrogate the model to extract what is driving the prediction. Another way to do this is to create a proxy to illustrate what the model is doing, such as a decision tree, which are easily viewed and understood by humans. The tool also

[12] Founded in 1956, FICO provides analytics software and tools used across multiple industries to manage risk, fight fraud, build more profitable customer relationships, optimize operations and meet strict government regulations. Their product FICO® Score, reached industry-wide adoption in the United States where it has become the standard measure of consumer credit risk.

[13] The term "data wrangling" refers to transforming and mapping data from one raw data form into another format to make it more appropriate and valuable for analytical purposes.

allows the data scientist to look at the patterns, the variable importance, rank ordering, etc., as well as enabling users to recognize complex patterns with a vast array of analytic techniques and algorithms, resulting in diagnostics that explain how the model is working.

FICO also looks at explainability through the lens of the business user (i.e. person in charge of the P&L for a product or portfolio). Typically, scorecards have been used which are easy to explain, but ML models have a very different model structure which is much more difficult to explain. FICO found that explaining a decision (as opposed to a model) to a business user requires the support of graphs, charts and statistics, and for that reason they created algorithms that simulate the different decisions that would be made based upon the model. The focus (which parallels our discussion in Section 2) has been to focus more on how to use the machine in the context of larger environment, rather than focus on the machinery in isolation.

FICO has found that ML models can add value at every step of the path to get to model, and not just in creating the models themselves, as outlined in Section 2.3. This case clearly illustrates the use of ML to automate data enrichment. For instance, if you have millions of transactional data records, FICO has a way to automate the generation of signals in the data, creating features that would normally be created in a manual process performed by a human expert. By shortening the time spent on data wrangling and enrichment, FICO clients are able to explore a much wider variety of data than before, resulting in improved human-technology collaboration. We also observed this in the IIF ML-CR report: the challenge of data (and IT infrastructure). With ML, firms can confidently transform the raw data to the right structure and volume necessary for modeling, by using a process which allows the signals to naturally surface without imposing too much judgement, which could bias the results.

Interpretability starts very early on in the process. When dealing with raw data, FICO enriches it methodically. They look at the features that are being created and gain insights into which features are having strong signals. The end-goal for FICO is to do a large volume of feature generation or characteristic generation in an automated way. The thinking is that even for the diagnostics coming out of the back of the model, such as variable importance, firms still need to understand and articulate what variable X is. The way they tackle this is by making feature generation transparent. In the financial services industry, they typically see transactional data in a variety of forms (e.g. credit, debit card transactions, website and mobile app data etc.), and generate characteristics from that transactional data by taking a templatized approach. A very basic example of templates might include one in which a template sums the count of transactions (e.g. the number of log-ons to the mobile app, or the number of credit card transactions) over a specific period of time (e.g. the last 7 days). FICO indicated that they have created a wide set of templates that comprehensively generate characteristics describing all the possible different trajectories of the transactions for a given customer.

Figure 5: Streamlined Process and Best in Class Tools[14]

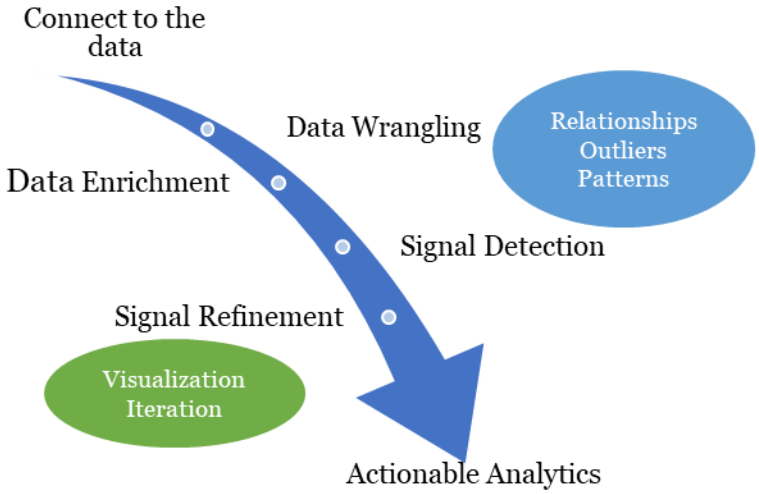


Figure 5 shows this streamlined process. For feature generation, they use different sets of techniques, which include the technique we describe above, for transactional data (anything with a timestamp), text analytics, and the use of network analysis. With network analysis, they construct the networks of entities and automatically generate a large number of characteristics which are aggregates (e.g. sums, averages, etc.) of the characteristics which describe the entities within a network, i.e. those entities associated with the entity of interest.

The process also involves visualizing those features (signals) that were extracted, for example, by correlating them with some of the behaviors of interest and collaborating with the human expert to understand those signals and patterns and iterating as needed.

Another way in which they show the impact of the signals on decisions have been to simulate the impact of the decision tree using an alternative score, and then seeing if the decisions themselves would change dramatically and why. In other cases, a client may ask them to drill down into a swap set- that is, to show the volume and profiles of persons the ML model would accept and the "classical" model would reject, and vice versa. This allows clients to understand more about the ML model and associated strategy, for example, by understanding how the types of accepted applicants (or applicants receiving a specific credit limit) would differ from their business as usual "classical" strategy.

In terms of understanding the drivers of the decisions (e.g. the typical reason codes and which variables contributed most to a particular decision), even if the score wasn't generated by a scorecard, users can still generate reason codes to help them understand a specific instance of a decision.

FICO is tackling explainability with a variety of methods to ensure that transparency is maintained from the raw data, and throughout the process to ensure that their whole team, including business users, can understand the outcome of the model.

[14] Source: FICO, Fair Isaac Corporation, Confidential, 2018

Conclusions

As firms progress in their adoption and application of Machine Learning techniques, they have proceeded cautiously, with intensive scrutiny applied across model testing, validation and the broader model risk management and governance process. As was highlighted in the *IIF ML-CR Report*, Machine Learning initiatives have attracted greater curiosity within firms' management, and internal governance often means that management are asking the same questions that a supervisor or regulator might.

Concurrently, the desire for Machine Learning models to be explained and transparent needs to be seen in context. As outlined in Section 2, not all existing linear or rule-based models can be considered as intrinsically transparent or meeting with an intuitive explanation; even when the weights of a linear model might seem intuitive, they can be fragile in terms of feature selection. It is one thing to expect Machine Learning techniques to demonstrate comparable interpretability as other (eg. linear regression) models, but they don't need to be held to an even higher standard.

In those cases where there is a demonstrable need for interpretability, banks and insurers are consequently facing up to the explainability challenge, incorporating several contemporary approaches to increase transparency and accountability. There is further opportunity to expand such approaches across the industry, though this needs several measures and approaches, and not relying on a single panacea.

It is important to acknowledge that not every algorithm needs to be explained, or its details understood – this depends entirely on the use case. It was notable (but not surprising) that while the IIF survey in credit risk applications identified a high focus on the explainability challenge, this was a comparatively smaller consideration for AML applications.

For instance, where the analytical capabilities of Machine Learning techniques are used for the creation of tailored attributes to be used in strategies or models, there is a need to be able to describe the meaning of the feature and explain their value compared to more traditional attributes. This can be done graphically, and we see firms doing it through residual analysis, or localized partial dependency plots.

On other aspects, as Machine Learning models expand into more areas, developers will be pressed to produce comparisons with existing models. For simple models, a common comparison is the weight of certain attributes in the model and their contribution to the risk assessment. For more complex and opaque models, other tools such as partial dependency plots could be used to make comparisons.

It is also increasingly important to determine what is meant by "interpretability." Without a common understanding of the definition, there are several conceivable possibilities of minimum attributes to be fulfilled. Therefore, we suggest a closer cooperation between regulators, supervisors, and the financial sector, keeping in mind the differing perspectives from a legal, technical, and policy-related point of view need to be aligned.

Finally, the use of Machine Learning is not siloed to the financial industry; in fact it is used successfully in many other industries, as well as competitors within the industry. Accordingly, we believe that regulators should welcome the use of new technologies so as not to stall innovation in the financial industry.

Resolving the remaining issues around explainability in a way that ensures these techniques remain usable and efficient is paramount. Financial institutions would welcome a stronger statement of support for the application of Machine Learning in their risk modelling processes, as well as more clarity on the expectation supervisors have regarding the methodology used. These techniques will not fundamentally change the way institutions build models for risk management purposes, but are a powerful tool to gain greater efficiency, better risk management capabilities and predictive accuracy.

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

Natalia Bailey
Associate Policy Advisor
nbailey@iif.com

References

Breiman, L. (2001). Random Forests . *Machine Learning, 45*(1), 5–32.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232. doi:10.1214/aos/1013203451. *The Annals of Statistics, 29*(5), 1189-1232. Retrieved from https://projecteuclid.org/euclid.aos/1013203451

IIF McKinsey. (2017). *Future of Risk Management in the Digital Era*.

Institute of International Finance (IIF). (March 2018). *Machine Learning in Credit Risk*.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. Retrieved from https://arxiv.org/abs/1606.03490

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

Ribeiro , M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. San Francisco, CA. Retrieved from https://arxiv.org/pdf/1602.04938.pdf

Shmueli, G. (2010). To Explain or to Predict. *Statistical Science, 25*(3), 289-310.

U.S. Department of Defense, Advanced Research Projects Agency. (2017). Explainable Artificial Intelligence. *DARPA/I2O Program Update*. Retrieved from U.S. Department of Defense.

Varian, H. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspectives 28:2*.