Tech Report 2021-01 , 8 March 2021, Version 1.0, ©2021 MRMIA

MACHINE LEARNING AND MODEL RISK MANAGEMENT

WORKGROUP MEMBERS

Peter Quell, Chairman
Anthony Graham Bellotti
Joseph L. Breeden
Javier Calvo Martin*

* On behalf of Management Solutions

Summary

An increasing number of business decisions within the financial industry are made in whole or in part by machine learning applications. Since the application of these approaches in business decisions implies various forms of model risks, this white paper tries to answer questions such as

- What are the essentials of machine learning? What are the corresponding dangers?
- Where lies model risk when using machine learning?
- How can banks mitigate these risks? What does a suitable validation framework look like?
- What do regulatory frameworks say about machine learning? Where does the industry need more clarification?
- Are there already established processes that could form the basis for the emergence of industry best practices when dealing with machine learning applications?

The following chapters offer some arguments to spark discussion among risk professionals. The working group for machine learning and model risk within the Model Risk Managers' International Association welcomes feedback on these issues under admin@mrmia.org.

MRMIA MODEL RISK MANAGERS' INTERNATIONAL ASSOCIATION

Table of Contents

Summary....1
1. Introduction to MRMIA....4
2. Introduction to the Machine Learning Workgroup....4
3. General Aspects of Machine Learning....5
3.1. Types of machine learning algorithms....5
3.2. Machine learning and statistics....7
4. Risk Model Validation....8
4.1. Which kind of application?....9
4.2. What aspects should be validated? And how?....11
4.2.1. Model data....11
4.2.2. Conceptual soundness....13
4.2.3. Model implementation and ongoing validation....16
4.2.4. Model documentation and use....17
4.3. Using Machine Learning to improve Validation....18
5. Model Risk Governance....20
5.1. The Regulatory View....20
5.2. A Governance Framework for Machine Learning....23
5.2.1. Model identification, registration, and planning of the lifecycle....23
5.2.2. Development, implementation, and use....25
5.2.3. Model review....27
6. Dangers of Machine Learning....29
6.1. Machine Learning and Explainability....29
6.2. Overfitting....30
6.3. Robustness and Population Drift....31
6.4. Bias, Adversarial Attacks, and Brittleness....32
6.5. Development Bias....33
6.6. p-Value Arbitrage....34
7. Risk specific Aspects....34
7.1 Credit Risk....34
7.1.1. What makes machine learning methods work in credit risk?....35
7.1.2. Role of unintended bias in credit risk....36
7.1.3. Adverse action notices....37
7.1.4. Imbalanced data sets....37
7.2 Market risk....38

8. Conclusions.....40
9. Members of the work group.....42

1. Introduction to MRMIA

Model Risk Managers' International Association (MRMIA.org) is a non-profit industry advocacy group focused on promoting best practices in model risk management. With a membership primarily from financial institutions, MRMIA seeks to provide an independent voice on model risk management that explicitly takes into account a cost-benefit analysis behind the practices being promoted. Specifically, MRMIA seeks to avoid chasing methods with diminishing returns, and instead seeks to promote those practices with real benefits to model developers, model owners, and those in various risk management functions who need to oversee the process.

This document was created by an MRMIA workgroup tasked with gathering best practices from personal experience, industry interviews, and academic publications. It is expected to be a living document that will be revised as new insights and methodologies become available. The views expressed here are those of MRMIA and the authors. They do not necessarily represent the views of their employers or any specific financial institution.

2. Introduction to the Machine Learning Workgroup

Machine learning has permeated almost all areas in which inferences are drawn from data. The range of applications in the financial industry spans from credit rating, loan approval processes in credit risk to automated trading, portfolio optimization, and scenario generation for market risk. Machine learning techniques can also be found in fraud prevention, anti-money laundering, efficiency / cost control and marketing models. Machine learning has demonstrated significant uplift in these business areas, and the use of machine learning will continue to be explored in the financial industry.

The banking industry is becoming increasingly aware of the model risks related to the use of machine learning techniques for risk management purposes. Even though quite comprehensive, regulatory guidance such as the Fed's SR 11-7 will not answer all financial practitioners' questions related to the implementation and use of machine learning algorithms in their daily business.

This white paper seeks to collect experiences from practitioners in the industry that could form the basis for a best practice approach for extending existing model risk and validation frameworks to machine learning applications. The workgroup and broader MRMIA

community have gathered these lessons and early thoughts about future impacts into the paper as a way to assist the industry and spur further industry research on these topics.

3. General Aspects of Machine Learning

Broadly speaking, machine learning is about computer algorithms that can learn to perform some task well. Typically, the task would require human intelligence to perform. Hence, machine learning is a sub-discipline of Artificial Intelligence.

Mitchell[1] describes machine learning in this way:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Here are some examples:

T (Task)	E (Experience)	P (Performance)
Detect spam email	Past history of emails	% spam detected correctly; % false alarms.
Credit risk management	Past credit data with delinquency and default information	Accuracy of credit default prediction and/or loss estimates.
Self-drive car	Past and on-going sensory data from car drives	Getting from A to B in good time, without crashing.

3.1. Types of machine learning algorithms

Experience may be presented to the learner in different ways, depending on the type of task. The three main categories of machine learning algorithms are supervised, unsupervised and reinforcement learning.

- **Supervised machine learning**

  A ***label*** is given for a data object, and the machine learning algorithm learns the association of the label to features of the object. The algorithm will learn based on the training data set of objects given with known labels. Once the learning is complete, the system can be used to predict the label for a new object. This

[1] Mitchell, T. (1990), Machine Learning (McGraw-Hill, 1st edition)

approach is called "supervised" since the learning is directed (supervised) according to the presented labels.

An example of a supervised machine learning problem would be predicting consumer credit risk based on a history of past consumer loans and their outcomes. The label here is a binary indicator of whether the loan was successfully repaid (0) or if default occurred (1). Traditionally, a linear model such as logistic regression can be used for this task, but alternative methods such as neural networks or gradient boosting can be used and may give improved performance[2].

Typically, ***predictive analytics*** tasks would be posed as supervised learning problems since they involve predicting labels.

- Unsupervised machine learning

  Unlabeled data objects are given to the machine learning algorithm to be arranged or clustered based on patterns to be discovered within the objects' features. Once trained, unsupervised machine learning algorithms can be used to sort objects into useful sub-groups. Typical unsupervised methods are clustering algorithms such as *k*-means clustering.

  An example of an unsupervised machine learning problem would be the discovery of different consumer types based on patterns in consumer transaction data. A machine learning algorithm capable of doing this would be useful for marketing.

- Reinforcement Learning

  The task to be learned occurs over time. Data objects are given unlabeled and the learner needs to perform actions across time. Depending on the consequence of the learner's action, the learner will receive feedback. Based on this feedback, the learner modifies their strategy (policy) and tries a new action. The process is repeated until the learner has learned a strategy that generally leads to good or useful actions for the task at hand. Typically, reinforcement learners are built on neural networks, such as Deep Q-Learning.

  Reinforcement Learning is the key machine learning technology for self-driving cars. In Finance, it can be used to support decision processes. For example, reinforcement

[2] Lessmann S., Baesens B., Seow H-V. and Thomas L.C. (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research*, Volume 247, Issue 1, 16 November 2015, Pages 124-136.

learning could be applied to the problem of developing an optimal policy for pricing of insurance or other financial products[3].

Machine learning algorithms come in two broad categories: parametric and nonparametric. A ***parametric*** machine learning algorithm is one based on some underlying model for which a fixed number of parameters need to be estimated to ensure optimal performance with respect to the task. Examples of this are neural networks or support vector machines. A ***nonparametric*** machine learning algorithm does not involve estimation of a fixed number of parameters and typically learns directly from training data. Examples of this are *k*-nearest neighbors and decision tree learners.

Many machine learning algorithms have hyperparameters that need to be set prior to their deployment and that control the learning process. For example, let's say we need to specify the learning rate of a neural network or the number of trees in a random forest. Even nonparametric algorithms use hyperparameters, such as the number of k neighbors in the k-nearest neighbors' algorithm. Often, the performance of the machine learning algorithm is highly sensitive to the choice of hyperparameters, hence care is required in their selection. Machine learning developers typically use grid search and cross-validation to achieve this careful selection, although more sophisticated methods, such as using genetic algorithms, is also possible.

3.2. Machine learning and statistics

There is a lively -- and sometimes heated -- debate about the distinction and overlap between machine learning and statistics. Machine learning has been developed in departments of computer science as a sub-field of Artificial Intelligence, itself a sub-field of cognitive science. This has occurred without a great deal of interaction with statisticians, so different techniques and jargon have been used -- often for very similar concepts. Given the broad definition of machine learning by Mitchell (cf. above), it could be argued that statistical inference is a sub-field of machine learning since inference can be seen as a learning task. However, many statisticians would argue with this. Indeed, some would claim that machine learning is actually a sub-discipline of statistics, since it is about making inferences on data and is assessed within a statistical framework.

Here are some characterizations[4] that distinguish machine learning from statistics. They are all broad generalizations, and exceptions can easily be found for each of these points. For further discussions see Carmichael et al[5].

[3] Krasheninnikova E., García J., Maestre R. and Fernández F. (2019), Reinforcement learning for pricing strategy optimization in the insurance industry, *Engineering Applications of Artificial Intelligence*, Volume 80, April 2019, Pages 8-19

[4] Bzdok, D., Altman, N., Krzywinski, M. (2018) Statistics versus machine learning, https://www.nature.com/articles/nmeth.4642.pdf?origin=ppub

[5] Carmichael, I., Marron, J.S. (2018) Data Science vs statistics: Two cultures? https://link.springer.com/article/10.1007/s42081-018-0009-3

1. Machine learning is about building optimizers based on objective functions that express the learning task.
2. Machine learning can handle high-dimensional and small sample size data in a way that statistics traditionally cannot.
3. Machine learning algorithms typically allow for non-linear expressions of solutions, whereas statistical methods are traditionally based on linear models. The greater the non-linearity, the greater the complexity of the learning algorithm. This can be controlled by the machine learning developer.
4. Statistical models are an attempt to estimate parameters of some underlying distribution generating the population. Some machine learning researchers do not even think in terms of a "true" data generating distribution.
5. In addition to (4), statistical and econometric models are often structural models that embody some underlying domain knowledge, whereas machine learning models are often free of a priori knowledge and just "let the data speak for themselves".
6. It follows from (4) and (5) above that, typically, machine learning will discover associations between variables without consideration of causality between features. On the other hand, statistical models can often be used to explain relationships in the real world. They are more meaningful.
7. Typically, statistical models are interpretable in the sense that we can explain how they work, whereas machine learning algorithms do not need to be interpretable and often are just "black boxes". For example, the operations of and decisions made by neural networks are difficult to explain.
8. Statistical models typically have stronger distributional assumptions than machine learning algorithms. Many machine learning algorithms assume that data are independent and identically distributed (i.i.d.), but that is all.

4. Risk Model Validation

Machine learning algorithms[6] have been proven to outperform traditional models in terms of predictive power in many situations and, perhaps most importantly, to be able to digest vast amounts of data, often unstructured and coming from a variety of sources[7]. This section will discuss the challenges and potential benefits that machine learning algorithms have for the internal validation function in banks and other financial institutions.

4.1. Which kind of application?

Two important distinctions should first be made in order to understand the context and current usage of these algorithms in the risk arena. The first is the distinction between regulatory and non-regulatory risk models:

- **Regulatory risk models** refer to those models subject to regulation and to supervisory approval. Examples of these are AIRB models leading to RWA calculation; IFRS 9 and CECL credit risk parameters; stress testing models used for CCAR, ICAAP or EBA exercises; derivatives pricing models; etc.
- **Non-regulatory risk models** refer to all other models used solely for risk management purposes with no need for explicit supervisory approval. Examples include credit collections models; credit risk early-warning systems; pre-provision net revenues (PPNR) forecasting models (except when used for CCAR or ICAAP), RAROC and loan pricing models, etc.

This distinction may vary with each particular jurisdiction. For example, in the United States, anti-money laundering (AML) models and fraud prevention models are heavily regulated and supervised[8], while in Europe this is not the case.

An immediate consequence of this is that machine learning models are still not widely used for regulatory risk modeling, arguably because they pose significant challenges for internal validation and especially the supervisory approval process. Regulators may not have clear answers, thus extending response times and increasing the risk of rejection. This may explain why the industry still favors traditional approaches (like the ubiquitous logistic regression) over more sophisticated and potentially more beneficial machine learning techniques.

[6] Even if machine-learning algorithms include traditional techniques such as logistic regression, for the sake of briefness we will use the term "machine learning" here to refer only to advanced non-traditional machine learning techniques, such as neural networks, random forests, gradient boosting, bagging, etc.
[7] "Artificial intelligence and machine learning in financial services", Financial Stability Board, November 2017
[8] "Artificial intelligence and machine learning in financial services", Financial Stability Board, November 2017

Case study

An interesting case study of this is the usage of a gradient boosting model for retail consumer loans admission in one European bank. After conducting research and benchmarking different techniques, the bank realized this algorithm significantly increased the accuracy in predicting defaults in this portfolio. It did so in a consistent manner: both type I and type II errors were reduced, and a backtesting exercise one year later showed that the predictive power was sustained over new loans, so there was no evidence of overfitting.

However, presenting the supervisor with this model for AIRB approval would be problematic for the reasons outlined previously. Therefore, the bank decided to adopt a strategy to solve this issue: they limited the usage of the gradient boosting to the scoring of consumer loans strictly for admission purposes, and two months after each loan was granted it would be scored under the consumer loans behavioral scoring model, which was indeed a traditional logistic regression. The sub-portfolio of consumer loans with less than two months' age was rather immaterial in terms of exposure, so a permanent partial use (PPU) of the standardized method may be requested in the future.

Through this regulatory workaround, the bank currently benefits from the predictive power of an advanced machine learning technique for an essential process (credit admission). Beyond the anecdote, this case study is representative of the aversion of banks towards the usage of machine learning algorithms for regulatory risk models.

Another aspect is the purpose for which machine learning techniques are used in the risk modeling domain. These are:

- **Exploratory data analysis,** in the sense of defining the segmentation, selecting variables, designing new variables, performing univariate and multivariate analyses, detecting outliers, etc.
- **Modeling**, in the sense of the choice of the algorithm (e.g. logistic regression, support vector machine, random forest, etc.) and the actual estimation of the parameters of the model, be it supervised, unsupervised or reinforcement learning based.

Currently, regulatory risk models in banks are using machine learning techniques mostly for exploratory data analysis (rather than for the actual modeling for the reasons described above), and non-regulatory risk models are using them to a greater (and increasing) extent, though still in an experimental and rather timid manner[9].

With this state of the art in mind, implying that machine learning models are still at a very early stage in the risk modeling of banks as compared to other industries (e.g. big tech, telecom or e-commerce), we will now review the implications of machine learning models for the risk model internal validation (IV) function in banks. We will approach the question from two angles: (i) IV as a validator of machine learning models, and (ii) IV as a user of machine learning techniques.

4.2. What aspects should be validated? And how?

Internal Validation increasingly faces the situation of reviewing models that incorporate machine learning techniques, either in the exploratory data analysis process or as the core algorithm of the models.

This poses significant challenges that may be grouped according to the elements which IV is typically entrusted to review (either by SR 11-7, by other local regulation or by common practice). A brief discussion of these challenges follows, together with some approaches which IV units in banks are exploring to address.

4.2.1. Model data

When validating machine learning models, an initial hurdle is how to assess the data used for their development, including the transformations that are applied to them, either manually by model developers or automatically by machine learning techniques.

A first challenge is **data representativeness**. If data are gathered from a variety of sources, merged, transformed, and sometimes sampled and filtered using automated machine learning techniques, how can IV verify the resulting dataset is still representative of the population it intends to model?

In this case, traditional representativeness analyses are usually still valid. Regardless of the transformation process, IV focuses on the resulting dataset and performs classical hypothesis tests (Z-test, Kolmogorov-Smirnov, chi-square, Kruskal-Wallis, etc.) to check for representativeness vis-à-vis the population to which the model will be applied, and any deviations from representativeness ought to be justified by model developers. It may be overly complex (and probably not worth it from a cost-benefit perspective) for IV to assess every single step in the data transformation process when machine learning techniques are involved, and a result-oriented approach may be sufficient in the current state of the art.

[9] "Artificial intelligence: challenges for the financial sector", ACPR, December 2018

A second challenge involves **data traceability and data quality**. How can IV verify that each particular data point used in the model estimation process is correct and traces back to the source system with no errors in the course of the extract, transform and load (ETL) process, when data may be unstructured, sources may be diverse, and machine learning techniques may have been applied in this process?

The answer relating to the quality of the original data typically exceeds IV's remit, and relies on whether the source system is certified as a golden source[10] (e.g. by the Chief Data Officer and/or in the context of the BCBS 239 framework).

As for the assurance of the absence of errors in the ETL process, perhaps with machine learning techniques involved, a common approach is an audit-like data lineage control framework based on checkpoints. That is, a number of control checkpoints are established throughout the data transformation process, and simple comparison indicators are observed in each intermediate dataset (e.g. number of records, mean and variance of each variable at every checkpoint, etc.), together with the rationale provided by model developers about why and how records are transformed or filtered out. This is typically complemented with a sample-based traceability analysis of a number of individual records.

A third challenge, notably more difficult than the former, involves **feature engineering** and the construction of synthetic variables that will be an input for the model estimation. How can IV assess whether these synthetic variables are correctly built, pertinent and fit-for-purpose, especially when they have been created using machine learning techniques and may not necessarily have a business intuition of their own?

In this case, IV units are commonly focusing on two aspects. The first aspect is challenging the theoretical description of the algorithm that performs the feature engineering (when it is automated) as to whether it is conceptually sound and whether it has been correctly implemented (e.g. if a code library from a reputed source has been used).

The second aspect is a deep analysis of the actual engineered variables in the resulting dataset. This includes both qualitative questions, such as the business intuition of each variable or to what extent the engineered variable may be "contaminated" by the target variable (thus rendering it artificially predictive and flawed), and quantitative checks such as the (linear or nonlinear) correlation with the target variable, or the distribution (number and size) of the buckets in categorical variables.

[10] "Artificial intelligence: opportunities, risks and recommendations for the financial sector". Luxembourg: Commission de Surveillance du Secteur Financier, 2018

Case study

In a European bank, in a behavioral credit scoring model used for regulatory (IRB) purposes, a combination of an XG Boost and an F-Race algorithm was used for the initial variable selection process and the variable binning process.

The algorithms were reviewed and approved by IV, though after many iterations due to the difficulty in replicating them. Nevertheless, in the words of the IV team, the usage of these machine learning algorithms contributed to the accelerate and optimize these processes. (Note: to date, they still have not been presented to the ECB as part of any application package.

Further challenges in model data are related to **other exploratory data analysis techniques** applied e.g. for outlier detection, univariate and multivariate analyses leading to the exclusion of variables, records, or sampling, among others. How should IV address the review of these techniques when machine learning techniques are used?

Once again, the current industry answer tends to rely on an end-result approach rather than on a debugging-like deep understanding of every step of the process. Do identified outliers make sense, or would they be ruled out as outliers by other techniques or data sources as well? Does the resulting sample dataset pass the abovementioned representativeness tests, and is it sufficient in a classical sample size test analysis? Then IV will not commonly delve further but will rather devote its efforts to the model itself.

4.2.2. Conceptual soundness

The most challenging task of IV regarding machine learning models is arguably the review of their conceptual soundness (in the SR 11-7 sense of the term), which includes the assessment of the model design, assumptions, limitations, explainability, interpretability, and potential bias and overfitting, among others.

As for the **model design and algorithm selection**, how should IV effectively challenge the choice of the machine learning algorithm, including the hyperparameters? How can IV assess whether the choice is poor, and an alternative option would be better? What are the techniques to perform this analysis, and what are the criteria to reach a conclusion?

There are no simple answers to these questions, since there is still no industry standard or regulation stating which algorithms are best suited or how the hyperparameters should be selected. Further the academic literature is still scarce in the banking risk domain, and a certain degree of subjectivity is unavoidable.

In terms of criteria for challenging the selection of a particular algorithm, model performance (predictive power) is certainly an important driver. The academic literature and some industry papers[11] broadly coincide in the conclusion that machine learning algorithms (e.g. random forests, boosting, neural networks) generally outperform traditional models (e.g. logistic regressions, trees) when there are strong nonlinear relationships between the regressors and the target variable, whereas they show little to no improvement when the relationships are roughly linear.

However, model performance should not be the only criterion, and perhaps not even the most important one, when assessing the choice of an algorithm. Additional drivers should weigh in, such as interpretability of the model[12], feedback from the banking supervisor (for regulated models), cost and effort of implementing, maintaining and monitoring the model, expertise within the bank on this particular technique, availability of reputed code libraries, and academic underpinning – all of these ultimately leading to an overall cost-benefit assessment.

From a more pragmatic point of view, this assessment is commonly supported by the use of challenger models and by different sets of hyperparameters. Whenever IV has sufficient capacity, it commonly departs from the same raw or processed data as model developers, and instead builds a suite of challenger models which normally include different choices of machine learning algorithms, different selections of hyperparameters, but also (most commonly) classic alternatives to machine learning models such as logistic regressions. It is not uncommon that challenger models built by IV end up outperforming the champion models both in terms of predictive power and the other abovementioned drivers.

Case study

In a European bank, a perceptron neural network with one hidden layer for a shadow-rating model was developed and submitted to IV. The IV team in turn tried a challenger model which was simply a multinomial logistic regression based on a monotonic Box-Cox transformation that linearized the variables.

This model roughly matched the predictive power of the neural network but with a classic, simpler, and more interpretable approach, and was eventually selected as the algorithm of choice for further shadow rating models.

In cases like these, where the challenger model outperforms and actually becomes the champion model, a bedeviled and uncomfortable question arises: if the final model was

[11] See, for example, https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine learning-challenges-lessons-and-opportunities-in-credit-risk-modeling
[12] "General principles for the use of Artificial Intelligence in the financial sector", DNB, 2019

developed (or heavily influenced) by IV, how is the independence of IV ensured? To this day, there does not seem to be a satisfactory answer to this question.

Regarding the **model assumptions and limitations** of a machine learning model, how should IV assess that they have all been identified and properly documented[13] and that none were overlooked, leading to a potentially serious model risk?

In this case, the answer does not rely on data science techniques or on hypothesis tests, but rather on the expert knowledge and deep understanding by the IV team of machine learning models and of the intended use of the particular model under review. Underlying assumptions like the stability of a nonlinear relationship over time, the representativeness and sample size in each bucket, or the permanence in time of the product mix of a portfolio, may not be straightforward to identify, assess and document.

A special case may be worth some attention: the usage of **dynamic learning**, in the sense of models that are automatically (and not manually) re-calibrated and re-adjusted on a continuous basis with no human intervention such as in the cases of incremental machine learning or online machine learning models. While these techniques are still not often used in the risk management domain (they are more common, e.g. in online advertising models), they are being increasingly explored for some non-regulatory algorithms such as fraud prevention models.

This case poses difficult challenges. How can IV effectively challenge a model that is permanently changing, sometimes literally with every new observation? And what is a "material change" in this context? This question has not been much explored as of yet since it is rare that IV is actually faced with this kind of algorithm (exceptions include, e.g. some credit collections models in digital banks). Key topics may revolve around two axes: (i) a deep methodological assessment of the dynamic recalibration technique, including a sensitivity analysis to the incorporation of outliers and observations that significantly deviate from the historical population, and (ii) periodic checkpoints where the latest recalibrated version of the model is backtested and assessed for consistency and business reasonableness[14].

Perhaps the most difficult issue when validating machine learning models is the **explainability and interpretability of the model**. This topic is being addressed at large in the industry, and some tentative industry standards are beginning to emerge, including LIME, SHAP[15] and surrogate models, as explained in the section on dangers of machine learning in this white paper.

[13] "General principles for the use of Artificial Intelligence in the financial sector", DNB, 2019
[14] "Artificial intelligence: opportunities, risks and recommendations for the financial sector". Luxembourg: Commission de Surveillance du Secteur Financier, 2018
[15] Lundberg, S.M., Lee, S. (2017) A Unified Approach to Interpreting Model Predictions; Ribeiro, M.T., Singh, S., Guestrin C., (2016) Why Should I Trust You?" Explaining the Predictions of Any Classifier

These techniques are necessarily imperfect. For example, if a simpler surrogate model explains sufficiently well the behavior of a complex machine learning model, without a major loss in predictive power, why not use it instead? They are commonly complemented with both large-scale sensitivity analysis and are selected on a case-by-case scenario analyses in order to get a better understanding of the model behavior.

Suffice it to say that interpretability is arguably IV's main concern about machine learning models, to the extent that there are already tools in the market specifically designed to help explain machine learning models' results.

A last point of concern within the conceptual soundness analysis is the outcome analysis -- specifically, the prevention of **overfitting and bias** in the model. How can IV assess whether a complex machine learning model is overfitted, or if it may lead to any potentially discriminating behavior due to bias in the input data or in the algorithm?

Again, this topic is not new in the industry. More traditional models, like classification trees, are also subject to overfitting, and there are classical techniques to prevent this. As explained in the section on the dangers of machine learning, model developers typically check this with regularization techniques, k-fold cross-validation[16], and a strict control of the learning curve.

To this regard, IV performs its own analyses, normally with the same kind of techniques complemented by individual case-by-case analysis on a sample of representative data.

4.2.3. Model implementation and ongoing validation

While IV teams are commonly more focused on the analysis of the model and its input data, some of them are traditionally less interested in assessing the correct implementation of the model -- especially in those regions where this is not strictly required by regulation (unlike e.g. SR 11-7, which explicitly requires that effective "controls and testing" are in place "to ensure proper implementation of models").

In the particular case of machine learning algorithms, where the model may contain a complex structure and large sets of parameters (possibly a dedicated platform with proprietary code libraries which pose computational challenges) how can IV validate that the model is properly implemented and there are no substantial errors?

In some cases, it may not be feasible or realistic for IV to do a full independent replication of the code or to debug it line by line to check for its correctness. In those cases, an external check may be sufficient because IV constructs and provides a large dataset with some millions of observations, either real or synthetic, and asks both the model development team and the model implementation team to score them with their versions of the machine

[16] "Artificial intelligence: opportunities, risks and recommendations for the financial sector". Luxembourg: Commission de Surveillance du Secteur Financier, 2018

learning model, and then checks that the outcomes exactly match. This may be sufficient evidence of a correct implementation.

It is also a requirement by most regulators that models are periodically monitored to ensure they continue to be fit-for-purpose, and, if not, to take action otherwise. This is not only in terms of statistical performance of the overall model, but also regarding other topics such as the stability of the underlying population, the significance of each individual variable in the model, etc.

Normally the first line of defense (e.g. the model development team) is in charge of performing this ongoing model monitoring function, and the IV team receives the results as inputs (among others) for its periodic model validation activities.

IV's participation in model monitoring per se tends to be limited, beyond reviewing and challenging the KPIs it provides. In the case of machine learning models, IV will normally review the set of KPIs it receives, and potentially proposes additional indicators to be periodically monitored, such as overfitting prevention metrics, a performance analysis over specific subpopulations, or indicators oriented to the interpretation of the model (such as the marginal contribution of each regressor or a local linear analysis).

4.2.4. Model documentation and use

An area of increasing attention is model documentation[17]. It is hard to exaggerate the importance of an exhaustive, traceable, self-explanatory model documentation, that according to regulation and best practice should allow the replication of the model by a third party.

In the case of machine learning models, what level of depth should this documentation reach? Should it consider an external code library, which provides the code for e.g. a random forest as sufficient documentation, or should it go deeper and detail every step in the random forest training algorithm? Should it presume that the reader is familiar with machine learning models, or should it become or contain a didactic manual, possibly lengthy, for non-specialized readers?

The answer varies by bank and depends on their current documentation standards, model risk culture and general degree of sophistication vis-à-vis the modeling practice[18]. Generally, IV needs to pay special attention and perform a deep and thorough review of model documentation to ensure that clarity, interpretability, and traceability of the models are up to the highest standards.

In the case of model use, IV's mandate is normally focused on (i) reviewing and challenging the procedures through which the first line of defense ensures that models are used exclusively for the purpose for which they were designed and approved, and whether that is

[17] "Ethics guidelines for trustworthy AI", high-level expert group on artificial intelligence, April 2019
[18] "Model artificial intelligence governance framework" (second edition), 21.01.2020, PDPC Singapore

indeed the case and (ii) assessing the procedures through which users provide feedback on the models.

The particular case of machine learning models only further emphasizes the need for this assessment by IV. In some cases, ad hoc workshops are organized between model owners, developers, validators, and users, where the new (machine learning) model approach is discussed, and users are explicitly trained and encouraged to observe and promptly report to model developers on any anomalous behavior in the model. Other than this, IV's activity on model use is usually not significantly disrupted by the introduction of machine learning models.

Last, but definitely not least, a major hurdle when validating sophisticated machine learning models is how qualified the IV team members are to effectively challenge them. A deep understanding of these techniques is needed, beyond the traditional statistical and computer science knowledge[19].

This calls for specific training, the incorporation of specialists in this domain, and the use of external assistance to incorporate this expertise and best practice in the IV team that may otherwise fail to provide an appropriate understanding and challenge of the models.

4.3. Using Machine Learning to improve Validation?

Up to this point, the role of IV as validators of machine learning models has been addressed. But can IV benefit from machine learning techniques as users, too?

In the current state-of-the-art, there are at least three ways in which this is being explored by IV teams in major banks:

- Challenger models
- Automated model building
- Optimization of the model validation process

The first way has already been discussed to some extent and consists in the development of machine learning models to challenge the model being validated. The challenger model is not necessarily intended to replace the current model (as mentioned above, it may encounter supervisory difficulties), but it nonetheless makes sense to try machine learning challenger models, since they may reveal hidden relationships among regressors and non-linear behaviors that may be otherwise overlooked.

[19] ECB Supervisory Board member speech 2018, "The digitalisation of banking – supervisory implications." https://www.bankingsupervision.europa.eu/press/speeches/date/2018/html/ssm.sp180606.en.html

The second way consists in the utilization of tools that automatically produce a massive number of models based on the same input dataset, then selects the ones that seem like a best fit and offers them to the user for review. The concept of "best fit" may be based on many factors besides model performance, including compliance with a set of user restrictions, or a well-balanced model in terms of weights of presence of variables from different profiles.

Again, even if this selection of models will not necessarily yield a replacement for the current model, it certainly provides useful insights, such as an estimation of the maximum "predictability" capacity of the dataset (i.e., what is the highest performance that can ever be expected with this dataset, regardless of whether the model makes business sense?) and unexpected combinations of regressors that a person would probably not try.

Case study

In the case of one German bank, IV was presented a stress testing time-series based (ARIMAX) model for validation. The model made perfect business sense, used the expected macroeconomic regressors (e.g. GDP, unemployment rate), had reasonable predictive power (R^2), and complied with all the expected tests (Durbin-Watson, Dickey-Fuller, etc.).

The IV team then used an automated Python-based time-series model development tool provided by an external consulting firm to produce 2,000 challenger models based on the same data. The tool allowed the user to introduce many restrictions for a model to be considered suitable (e.g. signs of the estimators, minimum and maximum number of regressors, minimum and maximum relative weights of the estimators, adequate p-values and test results).

Since the number of combinations exploded, in the sense that millions of models were theoretically possible, the tool used evolutionary (genetic) algorithms to rapidly converge and produce only those models with the highest chances of being suitable.

As a result, IV found that there was room for improvement in the predictive power of the model, and there were indeed combinations of regressors with business sense that would help in this direction. This feedback provided inspiration and ultimately led the model developers to produce a significantly better model.

The third way is probably the most difficult and the least explored: the automation or semi-automation of the model validation process using machine learning techniques. This

includes ideas such as using machine learning algorithms for cross-validation[20], feature extraction, and backtesting, but also more daring concepts like the drafting of the model validation report using natural language processing.

In the current state of the art, some progress has been made in this direction, but these initiatives are still in their very early stages. They are mostly limited to the automatic production of KPIs (quite similar to model monitoring) -- but not necessarily using machine learning techniques -- or to the automated filling of the test results in the periodic model validation report, reducing the manual workload, but again with no real intervention of advanced machine learning techniques.

5. Model Risk Governance

5.1. The Regulatory View

When it comes to model risk assessment, the Federal Reserve supervisory letter 2011-7 (or "SR 11-7" for short) is probably the most important regulatory text. Since machine learning algorithms did not permeate the industry so much by 2011, the document does not spend too much time on the corresponding issues mentioned in the previous chapters. To what extent does the banking industry see a need to update, modify or extend the SR 11-7 document?

Let us start with the SR 11-7 **definition of models**:

> . . . [The] term model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates. A model consists of three components: an information input component, which delivers assumptions and data to the model; a processing component, which transforms inputs into estimates; and a reporting component, which translates the estimates into useful business information.
>
> . . . The definition of model also covers quantitative approaches whose inputs are partially or wholly qualitative or based on expert judgment, provided that the output is quantitative in nature.

[20] Glowacki, J. and Reichhoff, M., 2020. *Effective Model Validation Using Machine Learning*. [online] Milliman.com. Available at: <https://www.milliman.com/en/insight/2017/effective-model-validation-using-machine learning/>.

Experts in machine learning sometimes claim that they do not use any theories (neither statistical, economic nor financial theories) or elaborate assumptions at all, they just let the data speak for themselves. Does that mean the use of machine learning algorithms will not imply any model risk? Certainly not, because the range of statistical, economic or mathematical *theories* need to be extended to capture elements of machine learning.

Once machine learning techniques qualify as a model, the **definition of model risk**

> . . . which is the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports. Model risk can lead to financial loss, poor business and strategic decision-making, or damage to a bank's reputation . . .

can be transferred immediately. SR 11-7 then continues with aspects concerning the design, implementation and use of models. These issues only need minor modifications before being applicable to QRMs based on machine learning technology.

What about **explainability** of results based on machine learning algorithms?

> Developers should ensure that the components work as intended, are appropriate for the intended business purpose, and are conceptually sound and are mathematically and statistically correct. Comparison with alternative theories and approaches is a fundamental component of a sound modeling process.

This definitely calls for the specification of additional regulatory expectations. Surrogate modeling -- namely the use of simpler (also statistical) models to explain the results of machine learning approaches -- could support the necessary comparison with alternative theories. Moreover, methods like LIME and SHAP (see next chapter) provide valuable insights.

Due to the relevance of training data for machine learning algorithms, the model risk in selecting this data basis needs to be addressed along the following lines.

> Developers should be able to demonstrate that such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology.

As already mentioned, machine learning may be perceived as “theory-free”, i.e. tests for **distributional assumptions** that do not provide too much further value:

> The nature of testing and analysis will depend on the type of model and will be judged by different criteria depending on the context. For example, the appropriate statistical tests depend on specific distributional assumptions and the purpose of the model.

> Furthermore, in many cases statistical tests cannot unambiguously reject false hypotheses or accept true ones based on sample information. Different tests have different strengths and weaknesses under different conditions. Any single test is rarely sufficient, so banks should apply a variety of tests to develop a sound model.

Validation has already gained much prominence in the machine learning community. How does this resonate with the SR 11-7 view? The first major issue is **conceptual soundness**:

> A sound development process will produce documented evidence in support of all model choices, including the overall theoretical construction, key assumptions, data, and specific mathematical calculations . . . As part of model validation, those model aspects should be subjected to critical analysis by both evaluating the quality and extent of developmental evidence, conducting additional analysis, and testing as necessary. Comparison to alternative theories and approaches should be included. Key assumptions and the choice of variables should be assessed, with analysis of their impact on model outputs and particular focus on any potential limitations.

Since "model choice" could be a quite elusive term within a machine learning framework, the above-cited paragraph from SR 11-7 definitely needs additional clarification. Obviously, model choices are not only restricted to such features as, e.g. "supervised learning" vs "unsupervised learning". Explainability aspects may support the analysis of conceptual soundness, but a consensus within the industry is yet to form.

Ongoing monitoring of machine learning techniques is especially important if essential features change once new data is presented to the algorithm.

> Such monitoring confirms that the model is appropriately implemented and is being used and is performing as intended.

Do we need a full re-validation after integration of a pre-defined amount of new data? Alternatively, is a continuous monitoring of performance measures (e.g. the learning curve) already enough? Does a recalibration of a neural network count as a model change?

Good **documentation** is also essential for machine learning algorithms.

> Documentation takes time and effort, and model developers and users who know the models well may not appreciate its value. Banks should therefore provide incentives to produce effective and complete model documentation. Model developers should have responsibility during model development for thorough documentation, which should be kept up to date as the model and application environment changes.

In this context, the industry must identify best practices for documentation of the inner workings of machine learning algorithms.

So where does this leave us when upgrading the model risk management framework for machine learning? SR 11-7 is quite general such that many aspects carry over to machine learning methods. Nevertheless, there are some issues that need to be clarified in order to form an industry best practice for a successful model risk management in the context of machine learning.

5.2. A Governance Framework for Machine Learning

As it is the case with most discoveries and improvements that bear great advantages, the use of machine learning models in the banking sector entails risks that require a conscious prevention. So far, this is hardly new information. At least since the publication of the SR 11-7 by the Fed, extensive model risk management frameworks are being implemented in order to identify, measure and manage model risks appropriately. However, the application of self-learning, opaque machine learning algorithms imply risks that previously were not associated to the use of models or else believed to be widely controllable. In this context, the applicability of existing frameworks for managing model risks should be questioned.

5.2.1. Model identification, registration, and planning of the model lifecycle

The foundation of every model risk management framework is the correct identification of the models in scope and their classification according to the intensity of model risk management activities required for each of them. Regarding the correct classification of machine learning models as such, two main changes can be observed:

- **Model identification process**: On one hand, not all institutions have at their disposal an automatically updated model inventory, and so must rely on punctual registration and verification processes. Due to the change from a stable number of models to a fast changing, unstable amount of machine learning models with short time-to-market requirements, more iterative and automatized processes will be required.
- **Model definition**: On the other hand, the already often highly debated decision of whether certain algorithms should be considered a model becomes more complicated as machine learning models take less traditional forms. For example, chat-bots in the client service that propose certain products to customers based on their own criteria do not correspond at a first glance to the traditional idea of a model. As machine learning models increase to not only provide support to the decision of human analysts but to autonomously take that decision and directly communicate with clients, the SR 11-7 limitation to approaches with outputs that are quantitative in nature, might not cover all types of machine learning models. The

exclusion in the SR 11-7 of "more qualitative approaches [...] which should also be subject to a rigorous control process [outside the scope of the SR 11-7]" might not be sufficient. However, removing the criterion of the "quantitative output" would widen the application scope of the guideline too much. A possible solution would be to find another common criterion for machine learning methods with qualitative outcomes, such as the degree of human control or their importance in business decisions, based on which the current definition can be overwritten.

Once the model is inventoried, the activities and effort required throughout the model lifecycle stages are determined. Banks aim at classifying model types into groups based on similarities in order to leverage synergies throughout the lifecycle. In the past, the resulting grouping of similar models in an inventory often resulted in quite intuitive groups based on the type of risk the models addressed (e.g. credit risk, market risk) and the high-level model type (e.g. PD, LGD). However, the wide range of different emerging machine learning technologies with multiple different formulations, applications and data usage, might require a grouping based on different characteristics (e.g. to assign the team with the most expertise in one particular machine learning methodology for the model review). Institutions should therefore extend their current model risk classification (tiering) systems, which in most institutions are based on the two axes of 1) materiality on the model impact and 2) relative importance of the use of the model within the firm (e.g. regulatory), with additional attributes.

- Existing dimensions of materiality/exposure and relative importance: While the materiality of the exposure and the relative importance remain relevant criteria, effects related to the relative importance might be extended. While regulatory relevance (e.g. in terms of capital calculation of pillar 1 models) is still a relevant criterion, risks of non-compliance with further regulation -- previously deemed to be controllable – such as GDPR data requirements or the discrimination of minorities through model biases, are to be taken into account.

- Complexity of methodology and design: The complexity of the model design becomes more relevant than ever. For machine learning models -- which learn autonomously and for which the steering possibility for developers is the framework defined for the model (e.g. hyper-parameters and the objective function) -- new ways of comparing the model complexity have to be found. This can encompass the chosen methodologies (e.g. highly complex neural networks vs. linear regression-GLM models) that determine the level of interpretability or indicators of the level of transparency, such as the amount of hidden layers or the number of parameters.

- Data usage: Data drives the complexity of the ML methodology and thus the difficulty in assessing the model components. Influencing factors to be evaluated are the volume of required data or number of data features, the complexity of data structures (e.g. unlabeled, metadata), the quality of data (e.g. poorly labeled, low

quality or unstructured data) and whether there are variable interactions and transformations.

- **Output parameters**: A further decisive factor is whether the model in question is based on supervised machine learning (with delimited output parameters, e.g. the prediction of a property price) or based on unsupervised learning, in which there is no direct way to evaluate output accuracy (e.g. sentiment analysis, clusters, recommendations).
- **Model recalibration**: An institution might determine if the model in question is static or requires continuous recalibrations. Thereby, the complexity varies depending on whether a potential recalibration of the model would require an entire redevelopment, or if the initial model structure might be maintained and only retraining the model with recent data would be required.
- **Testing and monitoring**: The capacity to conduct effective challenge drives the model prioritization, including the availability of benchmark models, the availability of cross validation testing showing good performance, and parameter stability across the samples.

Taking all the previous aspects into account, a risk-level can be assigned to the machine learning model as it can for traditional models, based on which the intensity of further actions throughout the model lifecycle can be determined.

5.2.2. Development, implementation, and use

If the SR 11-7 is to be used as a basis of assessing an institution's model risk management framework, several of the instructions are applicable as well for machine learning models given their principle-based formulation. Requirements like the documentation of the design, theory and logic of the model, as well as the comparison with alternative theories and approaches as a fundamental part of the development process, are applicable to machine learning models -- though they might require a different evaluation of what can be considered a "thorough documentation of the model design".

Given the "black box" nature and the constant evolvement of machine learning models based on the data they are nourished with, an exact documentation of the functioning and underlying assumptions (to the degree the model can be replicated by a third party) might become impossible. As of today, there is little guidance defining the expectations of regulators. Publications by the De Nederlandsche Bank (DNB), the PDPC Singapore, or the High-Level Expert Group on Artificial Intelligence (AI HLEG), which was set up by the European Union to define guidelines for achieving trustworthy AI frameworks, seem to have in common that they emphasize the importance of explainability but admit this might only be possible to a certain degree for some AI methodologies:

- The DNB defines "Transparency" as "financial firms should be able to explain how they use AI in their business processes, and (where reasonably appropriate) how these applications function." The DNB focuses on the transparency about policies and decisions related to the use of machine learning methodologies, the documentation of the reasons for the choice of specific machine learning models, their limitations, and situations in which the use should be discontinued. In particular, decisions that lead to more complex, less explainable machine learning methodologies should be appropriately documented and approved. The DNB relates the efforts that should be made to explain model outcomes to their materiality.

- Even though the AI HLEG stresses that the model's explicability is crucial, it admits it is not always possible to define how a model output has been obtained. It suggests that "in those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate". The AI HLEG further relates the transparency closely to the data traceability, for which it requests the documentation of data gathering, labelling, and the algorithms used to the best possible standard.

- The PDPC Singapore acknowledges the difficulty of achieving explainability for certain machine learning models. As a potential solution, it proposes the documentation explain how the model results are used for decisions and on its repeatability, i.e. the ability to produce the same results, given the same scenario.

Based on the regulatory suggestions, three main conclusions for the development of machine learning models can be derived as follows:

- For machine learning models with "black box" character, for which a profound explanation of the input-output-relation is impossible, the focus should lie on the more explainable framework components of the entire model application such as data inputs and human decisions made related to the model and its outcomes.

- The effort of explaining the model should depend on its materiality.

- Explainability cannot be replaced - but trust in the model can be built through complementary practices, such as the repeatability of model results.

As indicated above, the data which is used to build, train, and apply the model becomes an increasingly critical aspect regarding the accuracy and explainability of machine learning models. While SR 11-7 already stresses rigorous assessment and documentation of data quality and relevance, the use of greater amounts of (often unstructured) data from several sources demands more precise advice on data requirements to prevent specific risks

related to machine learning. As an example, the use of heterogeneous datasets (i.e., from different reliable sources) that are as complete as possible (i.e., no attributes are removed) may ensure that biases in the data can be recognized and treated accordingly. Ethical components of data use should be considered, such as the suggestion by the AI HLEG about preferring public sector data over personal data as a general principle. While the definition of specific data requirements is not the responsibility of the SR 11-7, the interconnectivity with different data requirements (BCBS239, GDPR, etc.) might be emphasized more clearly.

5.2.3. Model review

Given the great numbers of machine learning models, the frequently changing model functioning, and the greater impacts of errors, automatized ongoing controls become more relevant. Main changes from the review framework of traditional models can be observed in the three domains of review frequency, validation stakeholders/responsibilities, and review content.

Frequency: For machine learning models, monitoring is the new validation. While the SR 11-7 proposes that "banks should conduct a periodic review -- at least annually but more frequently if warranted -- of each model to determine whether it is working as intended", more frequent reviews are required for machine learning models. The frequency of assessing whether the model works appropriately might be based on the following main indicators:

- observed changes in key input values, e.g. macro-economic indicators
- established KPIs, which might be centered around early warning signs regarding the data or functioning of the model such as potential biases or unachieved targets, as well as external events like regulatory, legal, and technological changes
- the business volume of the model, given that the frequent assessment will lead to results that are more volatile

Validation stakeholders: Given the connectivity of machine learning models with an increased number of stakeholders, regarding both the input data sources as well as the use of the model outputs and the additionally affected risk areas (like the violation of data protection regulation or ethical policies), the different areas which should be involved as controlling instances from the beginning of the development onward should encompass stakeholders from additional domains such as HR, Compliance, and Operational Risk.

Furthermore, the existing approval structure might require modifications. Those members having higher familiarity with machine learning models might be included in the committees in order to avoid the rejection of models due to the lack of understanding of the underlying functioning.

Validation content: The SR 11-7 requires "all model components, including input, processing, and reporting" be subject to validation. While the test of input and reporting components is feasible for most AI models, the assessment of the processing component might be less feasible for "black box" models. The following tests gain relevance in situations where only the inputs and outputs of a model are observable:

- Focus on the implementation: The SR 11-7 requests that "the computer code implementing the model should be subject to rigorous quality and change control procedures to ensure that the code is correct, that it cannot be altered except by approved parties, and that all changes are logged and can be audited". In this context, a stricter validation of the implementation in the production environment becomes more important, given that -- depending on the model type -- at some institutions models are currently approved based on the model design only, while the correct implementation is ensured at a later stage through UATs.
- Benchmarking: The comparison of the model results to classical models is recommended in order to understand the reason for any deviation.
- Assessment based on different data sets: The SR 11-7 advice on analyzing the "in-sample fit and model performance in holdout samples (data set aside and not used to estimate the original model)" is one of the main applicable tests for machine learning models.
- Assessment of specific cases: As in the case of traditional models, extreme situations should be assessed. This not only encompasses stress testing through extreme input data values, but also the analysis of specific cases in which the decision has been made in favor of an obligor that was exactly on the edge of being neglected.
- Backtesting: Common sensitivity analysis and backtesting methods like the mean squared error might become ineffective, given that there is no longer a straightforward relation between inputs and outputs. K-fold cross validation techniques might become more suitable in this context.
- Reporting component: The assessment of model outputs "as part of validation to verify that they are accurate, complete, and informative and that they contain appropriate indicators of model performance and limitations" is relevant regarding the detection of potential biases in the model outcomes. For example, the portfolio distribution should be monitored in this context.

Based on the aforementioned reflections, especially regarding the development and review of machine learning models, it can be concluded that most of the high-level principles of the SR 11-7 are applicable. As a general approach it seems that existing definitions must be amplified (e.g. model definition or the definition of risks related to the use of models in order

to account for new arising ethical risks), while requirements related to the actual processes (e.g. accountability of model owners) must be made more precise.

6. Dangers of Machine Learning

6.1. Machine Learning and Explainability

Many machine learning algorithms such as neural networks are "black boxes" that are difficult to interpret or explain, as mentioned above. In finance, however, an interpretation is often required. This may be a legal or regulatory requirement; it may also be an internal requirement as part of the model validation process.

There are three main approaches to explainability:

- **LIME** (Local Interpretable Model-agnostic Explanations) analysis provides local interpretation of the machine learning algorithm's behavior using a linear model within some local region of feature space
- **SHAP** (SHapley Additive exPlanations) is a way to measure and visualize the marginal contribution of each feature to the machine learning solution
- **Surrogate models** use simple models, say, rule-based or decision trees, to explain the broad behavior of a machine learning algorithm

Each of these have weaknesses: LIME and SHAP only give a partial view of the behavior of the machine learning algorithm across a local region or for one feature at a time, whereas surrogate modeling provides only an approximate, incomplete explanation.

Approximation methods such as LIME and SHAP may not be appropriate, especially for critical applications in finance where precise explanations are required. Some machine learning algorithms are intrinsically explainable. For example, decision tree builders such as ID3 or CART are a class of such algorithms. The decision tree that is constructed is both a non-linear model and is also immediately interpretable by a human analyst. More sophisticated models such as Deep ReLU networks can be interpreted exactly using equivalent sets of local linear models (LLM). With the appropriate toolkit, the analyst can determine how the network is behaving locally with as much precision as is required.

When considering explainability, it is worth considering carefully what is actually required[21]:

- A full understanding of how the predictive algorithm works, or

[21] Agus Sudjianto, William Knauth, Rahul Singh, Zebin Yang and Aijun Zhang (2020), Unwrapping the Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification, arxiv.org/pdf/2011.04041.pdf

- An explanation of how the algorithm makes a prediction or decision for one specific case at a time

The first problem is harder than the second and may be needed for model validation. However, the business may only require providing the second, e.g. being able to explain a lending decision to a specific customer.

6.2. Overfitting

Because of the complexity of machine learning algorithms, they are prone to overfit training data, i.e. they fit spurious random aspects of the training data, as well as structure that is true for the entire population. There are techniques to handle overfitting, e.g. ***regularization***, and all machine learning developers should be making use of these techniques.

However, just applying regularization is insufficient, and the performance of the machine learning solution should be measured to check for overfit. Very simply, performance on the training data can be measured against that of an independent test data set. If it is very much better, then there is very likely an overfitting problem and the machine learning developer should improve the machine learning algorithm to reduce the overfit, e.g. increase the regularization term, reduce model complexity and number of parameters, or increase sample size if possible.

The ***learning curve*** should also be used to demonstrate the performance of the learner against training data sample size. Essentially, the learning curve will show performance plotted against the amount of training data available. The performance can be measured both on training and test data. The primary objective is to see if there is enough data to train the learner (i.e., does performance converge to some upper bound) or if more data is required. Secondly, since overfit is related to sample size, it can also indicate how a machine learning algorithm is overfitting and can project how much data is required before overfitting becomes no longer relevant or gets to an acceptable level.

Many machine learning methods use performance on an out-of-sample data set to determine when to stop training the model. This approach to preventing overfitting is generally effective, but it carries a caveat. As a rule, the more often a data point is tested the less it can be considered out-of-sample. This was recognized several decades ago. In the case of hypothesis testing, the significance of the result should be adjusted based upon the number of tests conducted, as in the Holm–Bonferroni method.

When repeatedly testing scoring metrics or goodness-of-fit measures on an out-of-sample data set rather than hypothesis testing as above, the author is not aware of an equivalent adjustment, but the same principles apply. Simply stated, a good result with fewer out-of-sample tests is better than a slightly better result after many more tests. This needs to be considered when creating machine- learning models and when reviewing completed work.

One solution is to have a third and final out-of-sample test data set. Of course, this is only effective if it is used very few times.

These principles apply to both scoring models tested across hold out samples and to time series models tested on an out-of-time sample. Rerunning an out-of-time test repeatedly can result in "look-ahead bias" where the meta-parameter decisions are based upon the analyst's judgement of accuracy on data that was supposed to be out-of-sample. This problem is particularly acute when modeling a short time series relative to the cycle being studied.

6.3. Robustness and Population Drift

Robustness in the statistical sense means that a statistical model still gives reliable parameter estimates even with deviations from modeling assumptions. Since machine learning algorithms work with few distributional assumptions, they are typically robust in this sense.

However, there is also a sense of robustness against changes in the distribution of the population, or ***population drift*** (sometimes also called concept drift). For the same reason that complex learning structures are prone to overfitting random variation in data, we might expect they are also prone to overfitting special conditions related to populations at a certain time. Thus, they may be more vulnerable to changes in population over time, in contrast to simpler linear models. In financial applications such as credit risk, population drift is typical. As a result, caution is required when developing machine learning solutions. Sensitivity to population drift can be tested by simulation study (artificially modifying the distribution) or by backtesting over a long period of data. Again, if there is a problem, solutions such as regularization or simplifying the learning structure may help.

If a machine learning algorithm is not robust to population change it should not be deployed for stress testing, since it may be especially unrepresentative for extreme values that are expressed in a stress test scenario. In machine learning there has been recent interest in ***Domain Adaptation Methods*** to address this problem of building a machine learning solution in one domain but applying it in a different domain[22].

This article is being written during the depths of the COVID-19 recession. As soon as shelter-in-place orders were issued in countries around the world, we all knew the corresponding credit risk models would have a problem. All the algorithms discussed here are data-driven pattern recognition engines. When past patterns are not predictive of future behavior, the models will fail to predict the outcome. The impacts of the COVID-19 crisis on model risk

[22] Kouw W.M. and Loog M. (2019), An introduction to domain adaptation and transfer learning, https://arxiv.org/pdf/1812.11806.pdf

management are so widespread that MRMIA has dedicated a separate workgroup to capture those lessons and best practices[23].

In a model driven world, we cannot simply wait months or years for new data to arrive to allow us to retrain the models. Even models explicitly designed to be self-adapting cannot adapt when the outcome is unknown. Human judgment is required to create intuitive models of how behavior is shifting and what adjustments or overlays should be deployed to compensate. In such crises users can understand more easily where model weaknesses might lie, where the presumed sensitivities are no longer true, and what adjustments might compensate for the new situation.

6.4. Bias, Adversarial Attacks, and Brittleness

Many machine learning approaches such as Deep Learning work on large data sets. It may be difficult to know if a large body of data embodies any bias (e.g. more men than women in the sample, or some sub-sample), but if it does then this may be a problem. Rather like overfitting, the complexity of machine learning solutions may mean that the algorithm overfits towards one group or another. Post hoc, the machine learning solution may be tested for bias across different groupings such as gender, race, or religion, if this information is available. If a bias is discovered in the data that adversely affects the machine learning algorithm, this can be rectified by a reweighting or under- or over-sampling method.

In areas that use Deep Learning such as image processing or self-driving cars, researchers have found ways to fool machine learning algorithms with adversarial attacks. For example, change just a few pixels of an image to make a face recognition system fail to recognize a face, or add black sticky tape to a road sign to make a self-driving car speed illegally. This is a consequence of the fact that machine learning algorithms are specialized, associative algorithms, rather than general intelligences with meaningful understanding of the world around them. The consequences for finance are uncertain, but we could imagine, e.g. a fraudster fooling a machine learning system into providing credit with a careful selection of inputs.

The narrow nature of tasks learned by machine learning algorithms can mean the algorithms behave naively or amorally, without appreciating the broader context in which they are operating. In particular, the machine learning algorithm may behave abnormally in regions of feature space where it has not been properly trained. Here are some examples:

- A cleaning robot that knocks over a vase because it was never trained to recognize it -- but cleans up the mess afterwards

[23] MRMIA, "Impacts of the COVID-19 Crisis on Model Risk Management", MRMIA WP No. 2020-01, December, 2020.

- A self-driving car stuck in a circle formed by a solid white line and a dashed white line, because it has been trained never to cross such a line
- A face recognition system that works excellently on a given image data set but fails dramatically on facial images from a different source

Brittleness means machine learning systems that work apparently intelligently in many circumstances, but which can become dumb very quickly in some special cases. For financial institutions, the implications could be profoundly serious, leading to large financial loss and reputation risk.

6.5. Development Bias

This is the bias introduced in machine learning as part of the development process. Very simply, this bias is introduced whenever developers iteratively repeat the development cycle until they achieve a satisfactory result.

In traditional statistical modeling, a typical way this can happen is through *p-value hacking* -- developers continually adjust the model until they achieve a p-value below some presupposed level (say, 0.05). The adjustments can be:

- Changing the data sample taken to train the model
- Changing the structure of the model (e.g. adding or removing variables)
- Transforming input variables

There are some cases where iterating over model build as part of the development cycle is legitimate, such as if it is clear, post hoc, that the modeling assumptions or data transformations made were very wrong. However, such re-evaluation should be kept to a minimum and should not be the norm.

With machine learning, the complexity of the learning structure means there is a great deal more for the developer to consider when producing the best machine learning solution. For example, neural networks allow for any number of neurons, organized into different numbers of layers. Other algorithms such as support vector machines (SVM) have their own set of hyperparameters that need to be tuned. Development bias is introduced when the developers manually manipulate these structures and hyperparameters until they meet some target performance measure.

If the performance measure is made on the training data, this process will very likely lead to overfitting. If the performance measure is made on the test data, and iterated model build is performed, then the test data set will no longer be independent from the machine learning development and can no longer be relied upon as an accurate unbiased measure of the future performance of the machine learning solution.

The overall principle must be that the test set should remain as independent of the machine learning development process as possible. This means developers should maintain a hold-out data set (which is usual practice for credit scorecard developers) and a separate validation data set or a ***cross-validation*** procedure can be used for choosing structure and tuning the hyperparameters. In particular, ***grid search*** is a systematic approach to search the space of hyperparameters in machine learning.

6.6. p-Value Arbitrage

In comparing machine learning with traditional methods, the worst reason to choose a winner would be if they were being judged by different standards. For the most part, machine learning models are considered acceptable if they test well out-of-sample, provide a reasonable disparate impact analysis, and do not appear to be biased. For logistic regression, the list is a bit longer.

The most notable difference is the use of p-values to screen for insignificant factors in logistic regression models. Standard practice among model validators and auditors is to make sure all coefficients in the model are statistically significant according to the p-value, given a reasonably chosen threshold. The p-value is essentially measuring the distance from zero considering the estimation uncertainty. For binned variables where each bin has a corresponding coefficient, the appropriate interpretation is that "some" of the bins should have statistically significant coefficients. For example, if month-of-year was an input with one coefficient for each month, you would not delete June from the model if its coefficient was zero, as long as other months were significantly non-zero.

The American Statistical Society says p-values should not be used or interpreted as a screening tool for rejecting inputs from a model, and yet it is standard practice in credit risk modeling. By using a p-value criterion for screening variables in regression models but not in machine learning models, we are creating a p-value arbitrage situation. Model validators must avoid creating a situation where analysts inadvertently choose machine learning methods over regression methods just because of inconsistent evaluation standards by those in model risk management.

7. Risk specific Aspects

7.1. Credit Risk

Machine learning methods received early attention from researchers in loan performance, but adoption into operational contexts has been understandably cautious. The earliest experiments were primarily in fraud detection, credit scoring, corporate bankruptcy, and

default forecasting. As machine learning methods have matured, parallel efforts have occurred in the application of those techniques to adoption in areas of credit risk, resulting in a wide range of new applications.

7.1.1. What makes machine learning methods work in credit risk?

Many studies have been conducted comparing machine learning methods to statistical methods. Although winners are assigned, the reasons some methods win are rarely identified. In one short study focused upon identifying the reasons for such success[24], a few patterns became clear. Machine learning algorithms, depending upon which one is chosen, may provide the following benefits for credit scoring as compared to traditional statistical methods:

- identifying nonlinear relationships between explanatory and predicted factors
- adapting to non-normal distributions for the explanatory factors
- finding interaction effects between explanatory factors
- achieving more robust forecasts through ensemble approaches

In all the above items, a talented analyst with sufficient time may achieve the same result. In fact, analysts can incorporate heuristic knowledge from previous experiences or adjust their mental model of how the world works in order to make choices that make the model more robust, though the reason for the choice would not be apparent from the limited data available. However, for the analyst to achieve this success, the analyst needs experience and time to experiment. Perhaps the biggest advantage of machine learning is automating most of what a talented analyst could achieve so that better results can be achieved in the limited time available.

Much of model validation in credit risk has traditionally been about applying heuristic criteria learned by the industry overall. That form of applied intuition is not possible with many machine learning architectures, so an imbalance is created between validating statistical models and validating machine learning models. Whether this is good or bad for the final model production is not yet clear.

For all its promise, machine learning presents some unique challenges to application in credit risk. Unlike applications in speech recognition or image processing, accuracy alone is not sufficient in lending. FCRA[25] guidelines require that lenders not discriminate against protected classes and that consumers are offered explanations for denial of credit. Such concerns have dramatically slowed the adoption of machine learning, and with good reason.

[24] Breeden, J. L. "A Survey of Machine Learning in Credit Risk", May 2020, DOI: 10.13140/RG.2.2.14520.37121
[25] https://www.ftc.gov/enforcement/statutes/fair-credit-reporting-act

These and other valid concerns in model risk management must be addressed before the models can be widely adopted.

7.1.2. Role of unintended bias in credit risk

Machine learning has been in production for fraud detection longer than any other application in lending. Conversations with those involved at the beginning suggest that the earliest efforts did not have zip code as an input but were essentially zip code detection tools. Using or inferring zip codes in loan underwriting or pricing is called redlining and is prohibited in the US under FCRA. In fraud detection, no such prohibition exists, and one wonders why they did not just give it zip code to start with.

This story is useful only in the notion that given many other inputs, a sophisticated machine learning algorithm recreated the data it needed most. This is the greatest danger for using machine learning in credit risk. With linear methods, we generally feel safe in saying that no information on protected class status was given to the model, so the results are unbiased. Although a weak assertion at best, even this cannot be said of machine learning[26], especially when given alternate inputs. Big Data and sophisticated modeling approaches create significant unobserved risks of inequality and unfair treatment[27].

In credit risk, a machine learning algorithm might infer protected class status using social media data, credit card transactions, branch transactions, etc. One such example showed that the digital footprint of an online borrower was as predictive as FICO score, yet all those digital footprint data elements probably correlate to protected class status[28]. Excluding protected data is insufficient to assert that the final model's forecasts do not correlate to protected status. Simple linear correlation is the standard for discrimination.

A significant amount of research is being conducted on how to identify and mitigate disparate impacts from machine learning. Current methods can largely be grouped into two approaches. One group is modifying the input data to prevent models from finding biases. The second group modifies the learning algorithm to add constraints that would enforce fairness conditions.

The challenge with both approaches is the need to tag the data with information about protected class status. If we knew the demographic data for each account in the training data, one could trivially run correlations to prove that no bias exists after applying one of the above methods or others. Unfortunately, a linear mindset underlies these regulations. US

[26] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S. (2019) Certifying and removing disparate im- pact. In proceedings of the 21sh ACM SIGKDD international conference on knowledge discovery and data mining, pages 259–268, 2015; and Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. Iowa L. Rev., 105:1257.
[27] O'Neil, C. (2016) Weapons of math destruction: How big data increases in- equality and threatens democracy. Broadway Books
[28] Berg, T., Burg,V., Gombovi´c,A., Puri. M. (2018) On the rise of fintechs–credit scoring using digital footprints. Technical report, National Bureau of Economic Research.

lenders are not allowed to save data about race, gender, and other such information for anything except mortgages, so they lack the data necessary to prove that the models are performing fairly. Something will need to change here.

7.1.3. Adverse action notices

The US Equal Credit Opportunity Act (ECOA)[29], as implemented by Regulation B, and the Fair Credit Reporting Act (FCRA), require lenders to provide Adverse Action Notices when a consumer is denied credit. These notices are specifically intended to be both understandable by the consumer and actionable in the sense that the consumer can make improvements in their financial position in order to qualify in the future. Machine learning has many applications in credit risk, but when it is the primary underwriting tool, it must have good answers for consumers.

Unlike the previous discussion about global interpretability, providing reasons for specific decisions is an inherently local problem. Several methods exist for this, but it remains an important area of research, referred to as the quest for explainable AI. Section **Error! Reference source not found.** explains some of the methods currently in use.

7.1.4. Imbalanced data sets

Another problem that is more prevalent in credit risk than generic machine learning applications is the extreme imbalance between outcomes[30]. For example, in a commercial loan portfolio, defaults might occur for only 0.1% of accounts. This imbalance means that many machine learning algorithms will be happy to classify the non-defaults while largely ignoring the defaults, resulting in ever poorer performance where it is needed most[31].

Two main approaches have been explored to address the data imbalance problem. Brown and Mues[32] tested a range of machine learning methods across data sets with varying levels of imbalance in defaults to identify those methods best suited to modeling data sets with different default rates. One notable result was that traditional methods like logistic regression and linear discriminant analysis are robust to the degree of imbalance in the data, so this is largely a machine learning question.

---

[29] https://www.justice.gov/crt/equal-credit-opportunity-act-3

[30] Japkowicz, N., Stephen, S. (2004) The class imbalance problem: A systematic study. Intelligent data analysis, 6(5):429–449, 2002; and Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1):20–29.

[31] Vinciotti, V., Hand,D.J., (2009) Scorecard construction with unbalanced class sizes. 2003; and Kenneth Kennedy, Brian Mac Namee, and Sarah Jane Delany. Learning without default: A study of one-class classification and the low-default portfolio problem. In Irish Conference on Artificial Intelligence and Cognitive Science, pages 174–187. Springer.

[32] Brown, I., Mues, C. (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3):3446–3453

Others have pursued various strategies of modifying the training data to create more balance[33]: over-sample the smaller class, under-sample the larger class, apply weights to the training data, or generate synthetic data to augment the lesser class, as with SMOTE[34]. Overall, the results appear to show that adding to the under-represented class is most effective, with SMOTE being a commonly used approach. SMOTE is a random sampling along hyperplanes connecting pairs of points in the smaller class, a linear interpolation. Since this is using a simple model to generate data to feed into a more sophisticated model, it is no surprise that other methods have been proposed.

With any data manipulation approach, the analyst must remember that the underlying probabilities are being modified. The resulting model may be used for scoring but will require work in order to reintroduce predictions of probabilities. The simplistic approach of introducing a scalar to adjust for the over-sampling is risky, as the sampling will not be perfectly uniform across the feature space, so the probabilities likely will not be accurately recreated locally. This risk is acute relative to economic cycles.

7.2. Market Risk

Based on current experience, the usage of machine learning techniques within regulatory risk models for market risk in Basel's Pillar I seems to be quite limited. This can at least partially be explained by the fact that after the financial market crisis that began in 2007, regulators focused to a large extent on improving the frameworks for the "classical" approaches to market risk measurement like the fundamental review of the trading book (FRTB)[35].

Nevertheless, banks are already using machine learning strategies in their risk model validation processes. Though many of these developments were undertaken in connection with FRTB implementation projects, the corresponding improvements in the validation framework already show some benefits. Think, for example, about the profit and loss attribution test within the FRTB. In this context, banks need to control for inconsistencies between the hypothetical P&L and the risk theoretical P&L. Machine learning techniques have been used to spot patterns in the difference between these two P&Ls in order to improve risk modeling. Nevertheless, risk model validation units do not have to wait until full implementation of FRTB to benefit from these developments. With the same approaches they can establish classification schemes that support distinction between different types of backtesting outliers. If there is an outlier, validators are usually interested if this was due to unexpected market movements or due to inherent risk model deficiencies. Having enough

[33] Huang, Y., Hung, C., Jiau, H.C., (2006) Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 7(4):720–747; Marques, A.I., Garcıa, V., Sanchez, J.S. (2013) On the suitability of resampling techniques for the class imbalance problem in credit scoring. Journal of the Operational Research Society, 64(7):1060– 1070.
[34] Chawla, N.V, Japkowicz, A., Kotcz, A. (2004) Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1):1–6.
[35] https://www.bis.org/bcbs/publ/d457.htm

training data based on historical backtest outliers (and that is a crucial point), they could use machine learning classification schemes such as support vector machines (SVM) to characterize those outliers that should be used as starting points for potential risk model improvements.

Another important area of application is the construction of benchmark models or challenger models to probe the RWA approaches for market risk in Pillar I. It is a well-known issue that classical historical simulation approaches have difficulties adjusting to changes in market volatility. During phases like the fourth quarter in 2008 (Lehman default) and the first two quarters of 2020 (pandemic), banks reported many consecutive backtest outliers indicating slow adaptation properties of their risk models. Using techniques from sequential learning[36], these historical simulation approaches can easily be upgraded to yield better performance.

Machine learning approaches can also provide valuable support for the analysis of changes in value at risk numbers. Usually, several thousand risk factors could influence the value at risk. Machine learning techniques support the analyst by suggesting risk factor subsets that could be responsible for some (yet) implausible results of the market risk calculation.

On the other side, there are some features in market risk measurement that could impose challenges for current machine learning methods. The first issue relates to the low signal-to-noise ratio usually encountered in financial markets data. This aspect is closely related to the above-mentioned danger of overfitting machine learning algorithms. Due to the large noise component, the algorithm might learn the noise pattern, even though (at least in theory) there is nothing relevant to learn there. A clear indication of overfitting in this case is a good performance of the algorithm applied to the training data versus a deteriorating performance when applied to new data.

Another issue that could challenge a specific machine learning approach is the large demand of computational resources. In the context of market risk calculations, banks usually add market data from the current trading day, and the algorithm should immediately integrate that information without re-running the lengthy calibration process. This calls for the usage of so-called online learning algorithms in contrast to batch learning approaches. Especially in the context of time series analysis the sequential order of the data carries important information that should not be left out of the analysis.

For Pillar I market risk models the main area of application for machine learning algorithms resides in the context of risk model validation. The future developments around the internal model approach under the FRTB will show how far machine learning algorithms can be embedded in RWA calculations.

[36] Quell,P. and Meyer, C. (2020), Risk Model Validation, 3rd Edition, RiskBooks

8. Conclusions

This white paper shows that machine learning has already permeated the financial industry to a considerable extent. Some banks have already developed frameworks to deal with the model risks of machine learning applications, while other banks are still in the midst of soul searching for viable starting points. There definitely is a need to share emerging industry best practices and to develop a comprehensive framework to assess model risks in machine learning applications. The members of the working group for machine learning and model risk within the Model Risk Managers' International Association have collected some "top three" issues that need to be addressed. But this collection of issues is a subjective choice of the authors. We invite all risk professionals to share their views on model risk and machine learning under aimrm@mrmia.org.

Top three aspects in establishing an effective validation framework:

- **Team qualification**. The initial validation team needs to gear up with sufficient knowledge and hands-on expertise in machine learning techniques. A deep understanding is needed beyond the traditional statistical and computer science knowledge. This calls for specific training, the incorporation of specialists in this domain, and the use of external assistance to incorporate this expertise and best practice in the initial validation team. Otherwise, the team may fail to provide an appropriate understanding and challenge of the models.

- **End-to-end review**. The entire model validation framework needs to be revised and adapted to the fact that machine learning algorithms will sooner or later be submitted for review and approval. This will not only raise challenges in the conceptual soundness chapter, but also in data, implementation, monitoring, documentation, and use – so a full revisiting is in order.

- **More complex is not always better**. An appropriate balance needs to be found between model performance and all the other factors, e.g. interpretability, feedback from the supervisor, cost and effort of implementing, maintaining and monitoring the model, expertise within the bank, availability of reputed code libraries, academic underpinning, etc. To this regard, initial validation units need to avoid both an overly conservative and regulatory-only-driven position (that would discourage developers from even trying, with the perspective of a disheartening validation process) and a rather shallow and performance-oriented approach (where e.g. interpretability may be overlooked). Reaching this complex equilibrium requires a calm and insightful cost-benefit analysis.

Top three aspects for model risk governance:

- **Begin with existing model risk frameworks.** Even though machine learning introduces new challenges for model risk management, enhanced model risk frameworks should not start from scratch. Financial institutions and regulators have gained much experience in risk model validation in the last years that could serve as solid basis for model governance topics related to machine learning.

- **Consider the new role of data.** The new paradigm states that machine learning is model free, and everything depends only on the data. Though that may not be literally true, the more important role of data needs to be addressed within model risk governance frameworks. This white paper addresses issues related to various forms of bias, overfitting, population drift and regime changes. The new uses of machine learning in the financial industry will reveal many further challenges related to data.

- **Add new perspectives to your model inventory.** When it comes to model risk classification, machine learning will increase the relevance of ethical aspects due to data bias, explainability of model results, and the role of the recalibration process. To facilitate a comprehensive model risk management framework, these attributes need to be considered when filling the model inventory.

Top three aspects to mitigate the dangers of machine learning:

- As with traditional models, **excellent quality standards** for model development are a basic requirement. Additional checks for overfitting and sensitivity analysis to test for robustness should be emphasized for machine learning. Techniques such as regularization or feature reduction can be used to mitigate against these risks, if necessary. Also, to avoid reputation risk, tests for possible bias and discrimination should be used. Since bias is often a consequence of the distribution of the training data, reweighting or resampling of training data may help address these problems.

- The lack of **transparency** of many machine learning methods is the underlying cause of many risks in application in the financial domain. Therefore, deploying methods such as those given in Section 6.1 are valuable to address some of the risks identified in Section 7.1. Having some global sense of how a machine learning model is behaving can be valuable and insightful, even if it is only an approximation to the behavior.

- **Population drift** and sensitivity to abrupt changes in behavior are potentially a bigger risk for machine learning algorithms than traditional models due to their greater

complexity. For this reason, it is essential that machine learning systems are carefully monitored post-deployment using a well-considered set of KPIs that cover overall predictive performance, calibration, and bias/discrimination. Challenger models should be available to replace the machine learning algorithm if performance deteriorates significantly.

9. Members of the work group

Anthony Graham Bellotti

Dr. Anthony Bellotti is Associate Professor in the School of Computing at University of Nottingham Ningbo, China. He received his PhD in machine learning from Royal Holloway, University of London in 2006 and was Research Fellow in the Credit Research Centre at the University of Edinburgh from 2007 to 2010. He was senior lecturer at Imperial College London until 2019 where he taught quantitative methods in retail finance. His main areas of research are in machine learning and credit risk modelling.

Joseph L. Breeden

Dr. Breeden has been designing and deploying risk management systems for loan portfolios since 1996. He founded Prescient Models in 2011, which focuses on portfolio and loan-level forecasting solutions for pricing, account management, CCAR, and CECL. He co-founded Deep Future Analytics in 2013 as a CUSO to bring solutions to credit unions and community banks. He is member of the board of directors of Upgrade, a San Francisco-based FinTech, an Associate Editor for the Journal of Risk Model Validation and for the Journal of Credit Risk, and a founding board member of the Model Risk Management International Association (mrmia.org). Dr. Breeden received separate BS degrees in mathematics and physics in 1987 from Indiana University. He earned a Ph.D. in physics in 1991 from the University of Illinois studying real-world applications of chaos theory and genetic algorithms.

Javier Calvo Martin

Javier Calvo Martín is a partner at Management Solutions (MS). He currently co-leads MS' offices in Germany, the Netherlands and France. He is responsible for the relationship with the European Central Bank and the public sector and leads the model risk practice at MS.

During his career, he has conducted a number of projects at major institutions in Europe and the USA, especially focusing on model risk and on model development and validation (IRB, IFRS 9, AMA, stress testing, economic capital). He also leads MS' Research and Development function.

Peter Quell

Dr. Peter Quell is Head of the Portfolio Analytics Team for Market and Credit Risk in the Risk Controlling Unit of DZ BANK AG in Frankfurt. He is responsible for methodological aspects of Internal Risk Models, Economic Capital and Model Risk. Prior to joining DZ BANK AG Peter was Manager at d-fine GmbH where he dealt with various aspects of Risk Management Systems in the Banking Industry. He holds a MSc. in Mathematical Finance from Oxford University and a PhD in Mathematics. Peter is member of the editorial board of the Journal of Risk Model Validation and a founding board member of the Model Risk Management International Association (mrmia.org).