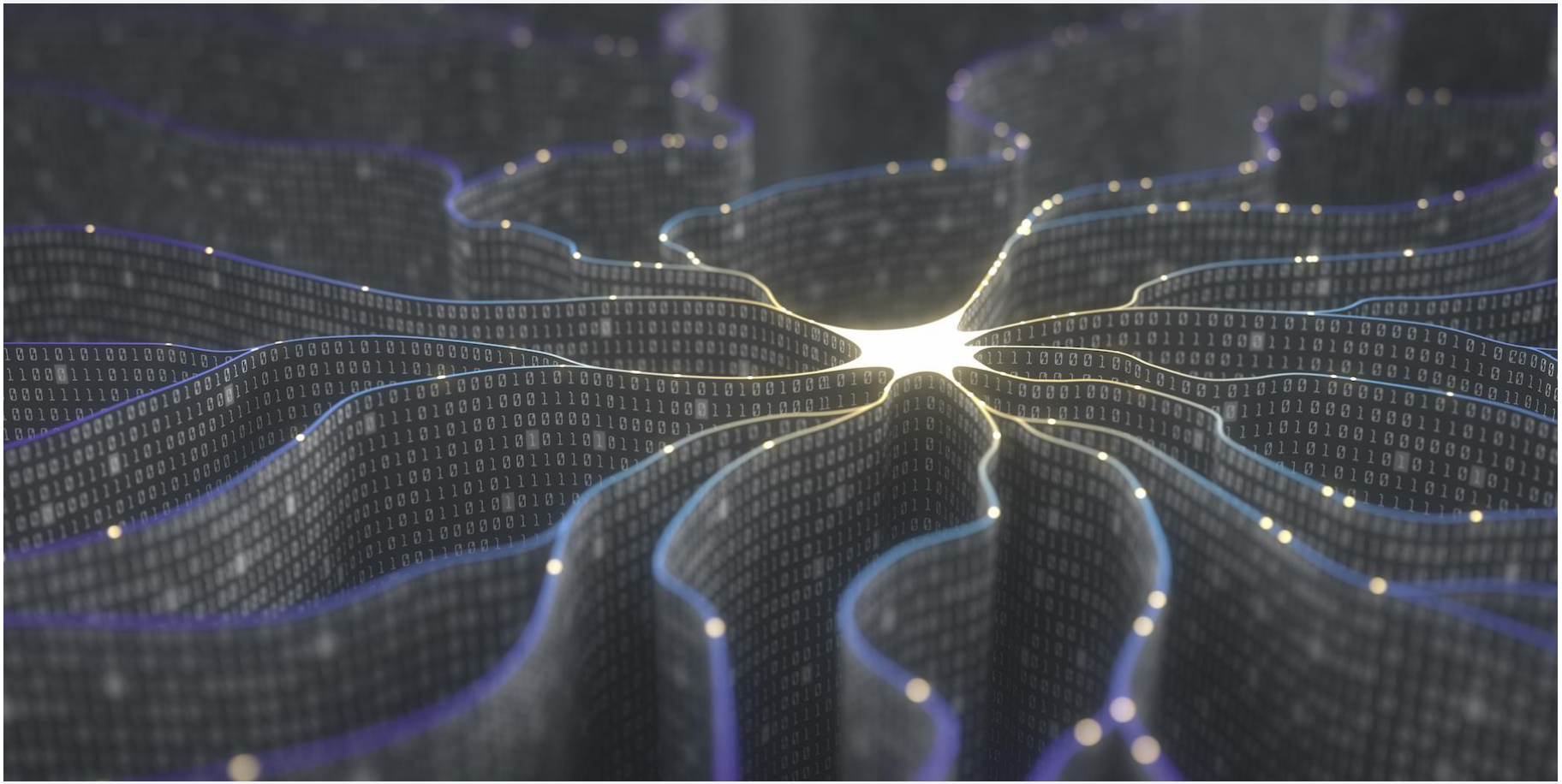
fintegral

3rd March 2021



Validation of Machine Learning Models

fintegral

Validation of Machine Learning Models

Introduction and selected aspects of validation

Modern Machine Learning techniques are becoming more ubiquitous and are finding widespread use in financial institutions. This opens questions about the treatment of these applications as models - including their validation. Below we briefly introduce the topic and summarise selected aspects to consider when validating machine learning-based models.

What is machine learning?

Machine Learning (ML): ML is a branch of artificial intelligence covering algorithms that learn through experience. In practice, ML algorithms rely on sample data (also known as training data) to make predictions. The term ML shall here be restricted to "modern" techniques, i.e., not considering regression as a ML technique.

Learning types: ML can be broadly divided into the following categories:

- Supervised learning derives the link between the model input and output based on example input-output pairs.
- Unsupervised learning finds patterns in unlabelled data (patterns such as clusters).
- Reinforcement learning uses a virtual agent that learns a strategy by itself based on feedback. It tries to maximise its cumulative reward, which is calculated using some scoring function.

Algorithms and applications: ML utilises a wide variety of algorithms (neural network, decision trees, k-nearest neighbours, support vector machines etc.) and covers an expanding range of applications. Examples are fraud prevention, anti-money laundering, chatbots, credit scoring and approval, and early warning systems.

Is a machine learning application a model?

Model definition: 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 (SR 11-7).

Model Risk Management (MRM) framework: Although there are some "academic" discussions, we are convinced that the definition above also covers the working principle of an ML algorithm. However, ML models have some specific characteristics that will need to be addressed in a comprehensive MRM framework.

Against this backdrop, banks should define a ML adoption strategy, which should be reflected in an updated MRM framework to address the specific aspects of ML. Exemplary tasks in this context are:

- Incorporate ML specifics in the model inventory and in the model tiering with respect to the subsequent model validation requirements.
- Include aspects like interpretability, bias/unfairness, and resilience to adversarial attacks in the content of the validation analysis as well as the model selection.

Selected aspects in the validation of ML models

Model selection and choice of ML algorithm

ML vs traditional: Validators (as well as developers) should evaluate and thoroughly justify the use of ML algorithms instead of traditional non-ML approaches. The challenge here is to find a trade-off especially between model accuracy, complexity, and transparency. For example, the extra accuracy produced by a random forest approach compared to a logistic regression needs to justify the increased complexity and reduced transparency.

ML complexity: Both the chosen ML learning type and the choice of ML algorithm should be assessed critically, including the number of parameters in the respective ML algorithm. When comparing two models, the model with less parameters is typically seen as the simpler one. Generally, one should apply Ockham's razor which is, given two models with a similar performance, the simpler one should be preferred. For example, a deep neural network has many more parameters than a decision-tree based model, making the decision-tree model the simpler one.

Conceptual soundness assessment

Conceptual soundness assessments analyse the quality of the model's fundamental design and construction. It is one of the main challenges with ML models due to their complexity.

Model parameters: ML algorithms can have (too) many internal parameters, which can lead to overfitting. This can be addressed by an adjustment of a fitting function (regularisation). However, a critical review of the regularisation is needed to ensure an optimised out-of-sample performance. An exemplary and often applied regularisation technique for deep neural networks is called dropout. It offers an efficient and effective regularisation that approximates training multiple networks with different architectures in parallel.

Representativeness and data leakage: Validators should check that the training data is representative for the data used in production and that no information is used in the training process that is not available in production. The latter is known as data leakage and it can occur in many different subtle ways. For example, transformations on the entire dataset before splitting the dataset into train and test sets can cause data leakage.

Validation of Machine Learning Models

Selected aspects of validation

Selected aspects in the validation of ML models (continued)

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision. The "black-box" nature of some ML algorithms obscures the origin of their decisions.

Importance of interpretability: Model interpretability plays a crucial role in business adoption, regulatory compliance, and human acceptance and trust. Without interpretability it is hard to test a model's outputs against business intuition. Lack of interpretability makes it difficult to judge how the model would respond to changes in the modelling conditions and environment, as well as when it would fail. Furthermore, certain ML use cases require a minimum degree of interpretability. For example, in the US, creditors must provide specific reasons for their decisions to applicants.

Interpretation tools: There are interpretation techniques to shed light on the characteristics that the trained ML model has learned. Such techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), help to explain the origins of individual predictions. These tools should be used by validators to get a better understanding of the model's logic and to analyse the model's edge case behaviour.

Interpretability example using LIME

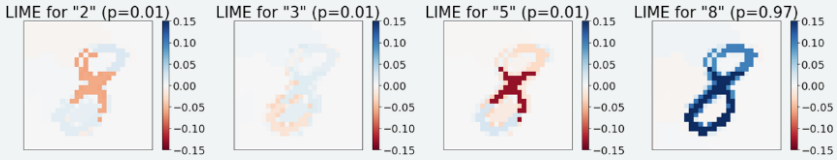
The ML model has been trained to recognise handwritten digits based on the MNIST handwritten digit database. The trained model attempts to recognise a previously unseen handwritten digit and calculates the probability p of the digit being 0 to 9:

Input

ML prediction

Digit	0	1	2	3	4	5	6	7	8	9
Probability p	0.0	0.0	0.01	0.01	0.0	0.01	0.0	0.0	0.97	0.0

We first select digits with probabilities $p \geq 0.01$. Using LIME we quantify the negative and positive contribution of sections of the image to the probability score. Below we visualise these sections and their corresponding amplitudes.



Example based on https://github.com/marcotcr/lime/blob/master/doc/notebooks/

Model bias & fairness

Model bias is when a model generates results that treats groups differently for no objective reason. An ML model is considered fair if its results are largely independent of given variables, especially those that are considered sensitive like gender, ethnicity, disability etc. Unfairness and bias are model weaknesses that might lead to legal and reputational risk.

Historical and sample biases: Biases can inadvertently get into a model in several ways. This includes through biases in the historical data that could reflect historical discrimination, or when the data are not representative of the population in question. Therefore, to reduce model bias, the training data used must be (i) representative for a wide range of groups and (ii) not badly affected by past discrimination. For example, a speech recognition model trained primarily on audio samples from women might produce biased outputs when used for male audio samples.

Fairness: An ML-based credit approval model which produces significantly different average approval rates for different ethnicities, could be unfair. Fairness can be evaluated on either model performance parameters (e.g., accuracy) or model output (such as an approval rate). Just leaving out the sensitive information in the training data is often not enough as other attributes are likely to be associated with sensitive variables. There are various techniques available to assess and correct for unfairness.

Ongoing monitoring of model performance

Ongoing monitoring of ML models should provide continuous insights into the model performance by using KPIs as well as other tools. In addition, monitoring mechanisms should be suitable to detect any anomalies in the end-to-end operation of the ML model and provide automated stop-loss-controls where necessary. Furthermore, the monitoring frequency should be risk-based, i.e., dependent on model use and materiality.

Other issues

In addition, there are many other aspects that should be investigated in model validation, regardless of whether one is dealing with a traditional model or an ML model. Examples are the sensitivity and robustness of the model, dependencies with respect to other models, implementation issues as well as the model infrastructure (data, processes, infrastructure, resources).

Contact

Fintegral

Frankfurt ¦ London ¦ Zurich

www.fintegral.com

Dr. Andreas Peter
Managing Partner
Fintegral Deutschland AG

+49 160 583 40 66
andreas.peter@fintegral.com

Fintegral Deutschland AG
Steinweg 5
60313 Frankfurt am Main
Germany

Dr. Pascal Böhi
Senior Manager
Fintegral Schweiz AG

+41 79 902 27 64
pascal.boehi@fintegral.com

Fintegral Schweiz AG
Brandschenkestrasse 150
8002 Zürich
Switzerland

Dr. Alexander Mottram
Senior Consultant
Fintegral UK Ltd.

+44 7703 788 016
alexander.mottram@fintegral.com

Fintegral UK Ltd.
City Tower, 40 Basinghall St.
London EC2V 5DE
United Kingdom