EMERTONDATA

Model Risk in the age of Artificial Intelligence and Machine Learning

By Aimé Lachapelle and Ali M'Rabeth

An **increasing reliance on Artificial Intelligence for decision making** is driving financial institutions, regulators, and supervisors towards a **clarification of sources and control of risks**. These risks were either already present (but marginal) or even non-existent in the usual model risk management framework. In a context where the use of machine learning is becoming massive and industrialized across banks and insurance companies, problematics such as **interpretability and dynamic monitoring, robustness, ethics, bias and fairness** require a specific attention.

Although all these topics are becoming active academic research topics and business innovation fields, their rigorous analysis from the ***model risk*** point of view remains at its early stage. A close collaboration between academics, regulator & supervisor experts and private sector professionals can accelerate finding pragmatic answers to multiple important questions, e.g. **how to interpret outputs of black-box models? How to monitor machine learning models in time?** When and why do they deviate? **How to control the discrimination** incurred by the algorithms? How to prevent the effects on decisions of input data changes or data falsification?

This short paper is based on Emerton Data research and analysis and provides **an introduction to the newly raised problematics of machine learning risks and ethics**, **with a focus on insurance** and more generally on financial services, probably the most mature sectors, even if these problematics will soon affect all industries.

1. Introduction

Financial services are by essence built around information and data flows. In this sector, quantitative decision making, notably based on data flows and data processing, has been a common practice for a while. There is a major change underway since **Machine Learning and Artificial Intelligence** are switching the financial services from a traditional modeling stage, where models were built by experts based on theories and assumptions, to a new modeling framework, where "black-box" models are trained to predict a required output based on inputs observations. **Therefore, models rely much less on experts' assumptions, and much more on input data**. To some extent, machine learning can be seen as an automation of model making, and the usual model risk is expected to be strongly shifted. Our objective here is to explore both the nature of the main model risk shifts and the degree of maturity of methods and tools to manage them.

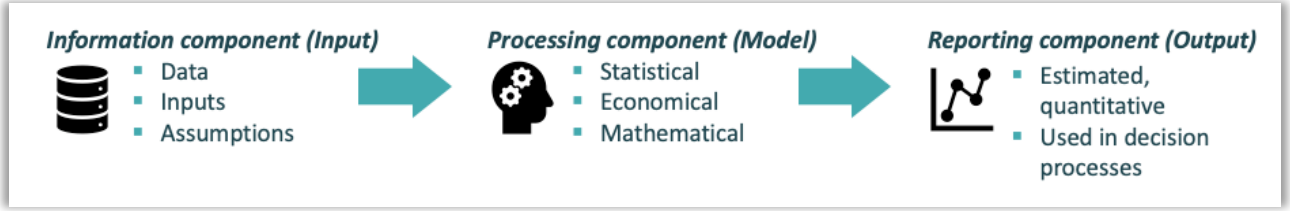
After clarifying definitions of model and model risk, we discuss why the use of machine learning has strong impacts on model risk, and we explore what we believe are the two major changes at stake: interpretability & dynamic monitoring on the one side, and fairness & bias on the other side.

2. Model Risk and Machine Learning

A **model** is a process that relies on statistical, financial, mathematical and economic techniques and theories, as well as on assumptions to operate input data into quantitative estimates for decision making (as illustrated in Exhibit 1).

A huge number of models operate for various purposes as for instance: to score credit risk, to price insurance policies, to define investment strategies, to improve CRM in marketing and communication, to decrease the costs and optimize processes in claims, to detect money laundering in financial institutions, etc.

Exhibit 1: What is a "model"?



Models are simplifications of reality; they are never perfect. Various metrics may be used to assess model quality, depending on its purpose. Performance metrics are well assessed. For example, in a fraud detection model, the objective might be to minimize the number of observations classified as non-fraudulent, while they are in fact frauds (also known as the false negative rate). However, many other model quality metrics can be assessed such as the robustness to some outliers, the stability to non-stationary inputs (fraud behavior do evolve in time), the non-discriminatory behavior, etc.

Model risk is a subset of operational risk, when a model is used to predict outputs (such as claims frequency, claims severity, elasticity) and crashes or performs poorly, leading to inadequate decision making (say pricing in the previous example) that can be seen as costs or losses.

It occurs primarily for two reasons:

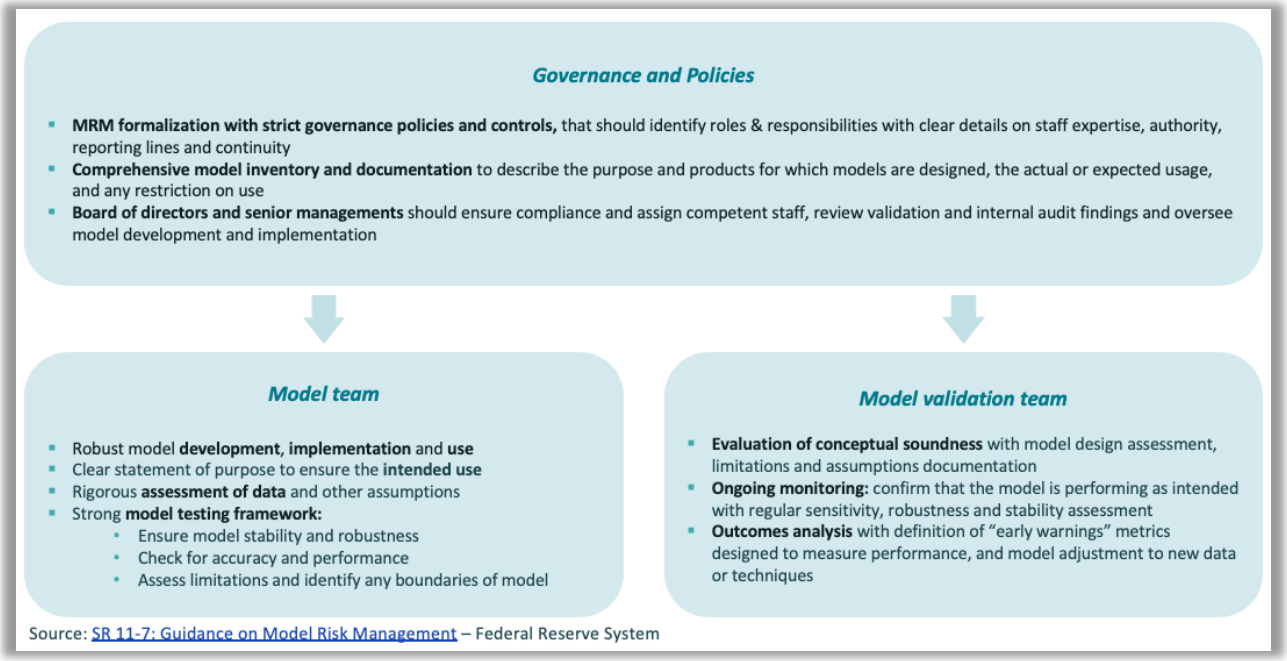
- *fundamental modeling errors* - models are based on theories and assumptions that are simplification of more complex phenomenon. Those approximations lead to a compromised outputs reliability and integrity. Moreover, the quality of inputs is determinant of the quality of a model: incorrect inputs, or non-representative inputs should lead to a defective model too,

and/or,

- *inappropriate model use* – a model is designed for a specific purpose, in a pre-defined environment. An application in a different environment from the one it was trained for may induce its inadequacy. Plus, calibration errors or data issues may have the same undesired effect.

The model risk can be addressed by an effective **Model Risk Management (MRM) framework**. The MRM framework suggested by the Federal Reserve is described in Exhibit 2. Its main objective is to guarantee the development and validation process across a whole institution. It embraces risk identification and assessment, risk measurement and mitigation, and risk monitoring and reporting.

Exhibit 2: Model Risk Management – Good Practices

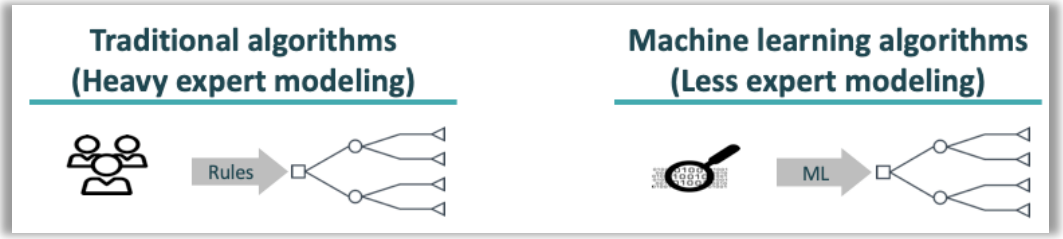


Source: SR 11-7: Guidance on Model Risk Management – Federal Reserve System

Financial institutions have been widely relying on models for a while. This dependency is growing with the rise of machine learning in behalf of the new applications that it enables, and the automation of traditional modeling tasks. **Therefore, the industrialized use of machine learning results in more models and less business modeling expertise.**

Indeed, machine learning models use less a priori assumptions, are less constrained, and need less expert modeling during the development phase (as illustrated in Exhibit 3). Their complexity enables a better performance by taking into account complex variables interactions such as multi-linearities or non-linearities. They can therefore contribute to reduce the risk associated with modeling, by detecting some correlations that humans might have underestimated or even not thought of. They will nevertheless generate new risks.

Exhibit 3: From traditional algorithms' design to machine learning algorithms



Let's take the canonical example of actuaries modeling risks in order to improve the price segmentation of insurance policies. These models are very critical since any underestimation of future claims costs can lead to lower prices. It can eventually lead to the insurer's bankruptcy because of adverse selection (many underpriced risks buy the policy). The traditional modeling task is time-consuming and based on many expert assumptions. Using machine learning could drastically speed-up the process and improve performance. However, how can one ensure that this machine learning model is robust (since there is no control on expert assumptions)? And we do not even mention the interpretability challenges (as often imposed by supervisors). The role of modeling cannot be overlooked. **An on-going and effective approach is to combine machine learning and business modeling.**

This is just one example (of the risk management challenges) among multiple use cases where nowadays machine learning is used in production: easy quote, fraud detection, claims settlement, litigation scores, deep triangle in reserving, marketing models, client targeting. For each of these applications, it is clear that new sources of risks emerge and that pre-existing risks (in the traditional modeling) are adjusted, leading to a necessary update of the model risk management framework. Rather than giving an exhaustive list of modified and new risks, **we propose to dive deeper on two major risks with machine learning models interpretability & dynamic monitoring, and fairness & bias.**

Interpretability and dynamic monitoring

Interpretability is the degree to which a human can understand the cause of a decision.

Interpretability and dynamic monitoring are amongst major challenges in machine learning. Stakes are at different scales. Regulators often require auditing the algorithms to ensure their compliance with rules. Internally, it is key (and even more in the AI adoption phase) that the organization understands the decision-making process, in order to audit and monitor decisions. Besides, insurers shall provide meaningful explanation of the decision making to a specific customer, and insights about how to improve their score for example. This is exactly the same for banks that need to justify to customers the reasons for credit rejection.

To address these challenges, there are three levels of interpretability: **algorithm transparency, global interpretability, and local interpretability**. They respectively answer to the following questions: "how does the algorithm create the model?", "how does the trained model make prediction?", and "why does the model make a certain prediction for a certain instance?". From a **global interpretability** perspective, it is possible to train an **interpretable surrogate model** to explain a complex "black-box" model. Such an approach is however limited due to performance issues. An alternative approach, with many recent applications, is to understand the decision-making process of a particular prediction, through **local interpretability.** For instance, if a client gets his/her credit request denied and asks how to improve his/her score, a financial institution can train a **local surrogate model***,* more interpretable, to understand how this score is computed and how to improve it. The most famous methods for such local interpretability are LIME and SHAP. Such methods go in the right direction, but they probably still lack efficiency to unlock a massive adoption.

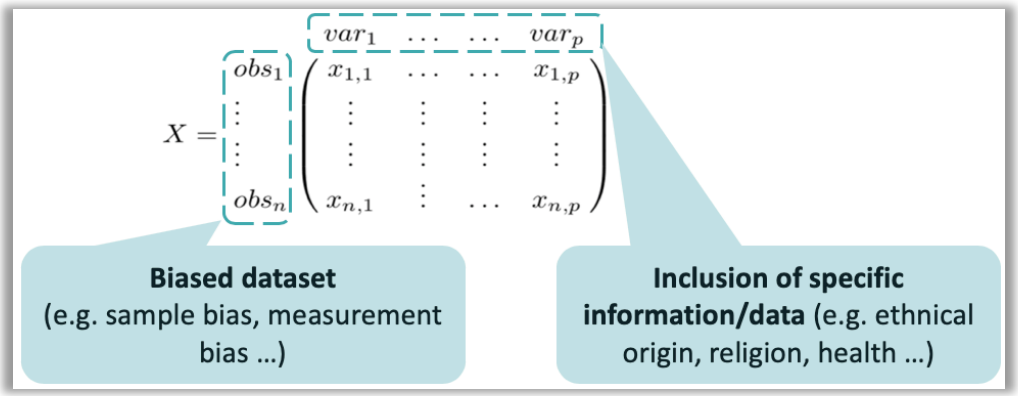
Last but not least, hundreds of models operate and the relation between inputs space and target space is barely invariant. Either the source generating the data is not stationary, and/or the concept to be learned changes over time, e.g. fraud behaviors evolve in time and might even adjust to fraud detection models in the run! Either way, the model should be robust to those shifts, by *detecting*, *understanding* and *dealing* with them appropriately. It is thus necessary to *understand* and *discern* this change, and not to modify the model despite a loss of performance. This is part of the ***model life cycle***: understanding when and why the model becomes obsolete, and how to fix it. Of course, this task is heavily based on interpretability. It is hard to understand how to fix models if you don't understand them.

Fairness and Bias

Dealing with **bias** and **fairness** is another major challenge in machine learning. In many use cases involving humans (e.g. credit scoring, insurance pricing, or even models taking into account employee's actions), models are expected to reach a certain level of fairness, in order to avoid any uncontrolled discrimination against groups of people. The *convention for the protection of human rights and Fundamental Freedoms* prohibits ''discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features […] age or sexual orientation''. Serving the objective to avoid discrimination can be completely antagonist to price segmentation for instance. There are two major sources of unfairness, as depicted in

Exhibit 4. For classical tabular data, **unfairness can basically come from the rows** (bias) **or from the columns** (information retrieval).

Exhibit 4: Two potential sources of unfairness in tabular data



The classical example in the insurance field is avoiding gender discrimination. By excluding the gender variable from a model, it is clear that all things equal, the output is the same. Nevertheless, this variable is usually correlated with others, thus the model will learn from those correlations (acting as *proxies* for the gender) ending up being discriminative for a gender as soon as one compares the models' outputs on a real set of females versus males. Therefore, a clear policy of fairness must be defined. Regulators understand that discrimination is not always a bad thing, and they are starting to differentiate between **intended discrimination** (prohibit the direct use of an information in models, e.g. gender, religion), and **discrimination as a result** (control and tolerate the observed shift of models' outputs by gender categories, etc.).

Two classes of approaches to address fairness are currently emerging in the academic literature (and to a less extent in banks and insurance companies).

- **"Fairness in the make",** which amounts to take into account equitability and prejudice criteria *at the time of designing and training machine learning models* (for instance by adding an equity term in the objective of the model),
- **"Fairness in the use",** which by opposition *rather focuses on the use of machine learning models' outputs* in order to satisfy to some equitability criteria (e.g. recruitment racial quotas in the US).

A **biased** training dataset is another main reason to unfairness. If a dataset is discriminative against a group of people, the machine learning model will reproduce this bias and can often even amplify it. We can cite here an example in the human resources field: if a recruiter unintentionally discriminates against a group of people, a machine learning model trained with data coming from his/her selection will reproduce/amplify this bias. One challenge is to learn to find out biases within a dataset and select which ones are harmful at a certain degree. Datasets certifications may exist in the near future.

Conclusion

Model risk has been identified and controlled in banks and financial services for a while. Since the sub-prime crisis, banks focused on strengthening their risk management framework, supported by new regulations and financial supervisions. **With the advent of machine learning in production, model risk is mutating. It is a growing concern and a special attention is being given to it.** Despite the recent works on new sources of risk, it is unclear that the financial services sector has reached the right level of maturity to properly assess and control machine learning risks. Interpretability of black-box models is well-identified but still poorly tackled. However, it is clear that interpretability is key for proper dynamic monitoring of hundreds of models (trained very frequently in many cases). Fairness is probably not well defined but is critical to control bias and discrimination that results from machine learning models.

At this stage the situation is clear: **insurers and bankers are probably not where they should be in terms of *model risk* in the age of AI and machine learning, and this could become an obstacle to further adoption of AI and machine learning in banks and insurance companies.**

What about other sectors? The industrial sector is also currently knowing a huge data & AI transformation period. Poor control of machine learning models has not been an issue so far, since the use cases involving machine learning had marginal impacts. This is progressively inverting, **so that all other industries will soon (if not now!) have to consider how to deal with *model risk* in the age of AI and machine learning,** although with slightly different priorities (fairness is indeed less meaningful when algorithms process machines' generated data). This topic will be addressed in a forthcoming paper.

Further readings on the topic

- **A reference note on model risk management** (finance supervision community): SR 11-7: Guidance on Model Risk Management, Federal Reserve System
- **A recent paper on the model risk fundamentals** (risk management community): B. Hassani, Model risk management: from epistemology to corporate governance, Journal of Risk Model Validation, 2019
- **A paper that gives some risk management perspective to AI and machine learning systems** (risk management community): Artificial Intelligence, Data, Ethics: An Holistic Approach for Risks and Regulation, A. Bogroff, D. Guegan, pre-print, 2019
- **The summary from the French regulator task force on AI challenges in the financial sector** (finance supervision community, in french): Intelligence Artificielle : enjeux pour le secteur financier, O. Fliche, S. Yang, 2018
- **The website of the AXA research department on these topics** (insurance company community, several papers on-line): AXA Rev Research website
- **A paper regarding "Fairness in the make"** (academic community): T. Kamishima, S. Akaho and J. Sakuma, Fairness-aware Learning through Regularization Approach, Proceedings - IEEE International Conference on Data Mining, 2011
- **A recent paper regarding "Fairness in the use"** (academic community): Kleinberg *et. al.,* Algorithmic Fairness, AEA Papers and Proceedings, 2018
- **The reference paper for LIME methods for interpretability** (academic community): M. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016
- **The reference paper for SHAP methods for interpretability** (academic community): S. M. Lundberg and S.I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems. 2017

ABOUT EMERTON DATA:

- **Emerton Data** is the dedicated Emerton entity for Artificial Intelligence and Advanced Data Analytics activities. Emerton Data supports organizations in designing and executing their data & AI transformation**. Visit http://www.emerton-data.com/**
- **Emerton** is a premier global, mid-size strategy consulting firm with offices in Europe, the Middle East and North America, blending strategy consultants and seasoned industry professionals
- **Contacts**:

  - **Aime Lachapelle, Partner:** aime.lachapelle@emerton-data.com
  - **Sebastien Plessis, Partner:** sebastien.plessis@emerton-data.com
  - **Pascal Simon, Partner:** pascal.simon@emerton-data.com

EMERTONDATA