June 2020

Governance of Artificial Intelligence in Finance

Discussion document

AUTHORS
Laurent Dupont, Olivier Fliche, Su Yang
Fintech-Innovation Hub, ACPR

Table of Contents

Front page: the CSIRAC, one of the first five computers put in service, supervised by its co-designer Trevor Pearcey (archival photograph - November 5, 1952).

1. Executive summary

This discussion document follows upon work led by the ACPR on Artificial Intelligence (AI) since 2018. In March 2019, after an initial report and a first public consultation, the ACPR launched, along with a few actors in the financial sector, exploratory works aiming to shed light on the issues of explainability and governance of AI – mainly understood as Machine Learning (ML). Composed of meetings and technical workshops, they covered three topics: anti-money laundering and combating the financing of terrorism (AML-CFT), internal models (specifically credit scoring), and customer protection. Two focal areas emerged from those works, namely evaluation and governance of AI algorithms.

Evaluation

Four interdependent criteria for evaluating AI algorithms and tools in finance were identified:

1. **Appropriate data management** is a fundamental issue for every algorithm, as both performance and regulatory compliance are conditional upon it. Ethical considerations, such as fairness of processing and the absence of discriminatory bias, have to be taken into account in this regard.
2. **Performance** of an ML algorithm can be addressed using a variety of metrics. The range of metrics available is sufficient for assessing the accuracy of virtually any ML algorithm used in finance, according to both technical and functional criteria. It is however sometimes necessary to balance the selected criteria against the desired degree of explainability.
3. **Stability** describes how robust and resilient an ML algorithm's behaviour turns out to be over its lifecycle. Due care must be taken to guarantee its generalizability to production data and to continuously monitor risks of model drift once deployed in production.
4. **Explainability**, a close cousin of algorithmic transparency and interpretability, has to be put in context in order to define its actual purpose. The "explanation" of a specific result or of the algorithm's behaviour may prove necessary for end users (whether customers or internal users); in other cases, it will serve those tasked with the compliance or governance of the algorithm. The provided explanation thus aims to either inform the customer, ensure the consistency of workflows wherein humans make decisions, or facilitate validation and monitoring of ML models. We therefore introduce four levels of explanation (observation, justification, approximation, and replication) in order to clarify the expectations in terms of explainability of AI in finance, depending on the targeted audience and the associated business risk.

Governance

Incorporating AI into business processes in finance inevitably impacts their governance. We recommend to particularly focus, as early as the algorithm's design phase, on the following aspects.

Integration into business processes. Does the AI component fulfil a critical function, by dint of its operational role or of the associated compliance risk? Does the engineering process follow a well-defined methodology throughout the ML lifecycle (from algorithmic design to monitoring in production), in the sense of reproducibility, quality assurance, architectural design, auditability, and automation?

Human/algorithm interactions. Those can require a particular kind of explainability, intended either for internal operators who need to confirm or reject an algorithm's output, or for customers who are entitled to understand the decisions impacting them or the commercial offers made to them. Besides, processes involving AI often leave room for human intervention, which is beneficial or even necessary, but also bears new risks. Such new risks include the introduction of biases into the explanation of an algorithm’s output, or a stronger feeling of engaging one’s responsibility when contradicting the algorithm than when confirming its decisions.

Security and outsourcing. ML models are exposed to new kinds of attacks. Furthermore strategies such as development outsourcing, skills outsourcing, and external hosting should undergo careful risk assessment. More generally, third-party risks should be evaluated.

Initial validation process. This process must often be re-examined when designing an AI algorithm intended for augmenting or altering an existing process. For instance, the governance framework applicable to a business line may in some cases be maintained, while in other cases it will have to be updated before putting the AI component into production.

Continuous validation process. The governance of an ML algorithm also presents challenges after its deployment in production. For example, its continuous monitoring requires technical expertise and ML-specific tools in order to ensure the aforementioned principles are followed over time (appropriate data management, predictive accuracy, stability, and availability of valid explanations).

Audit. As for the audit (both internal and external) of AI-based systems in finance, exploratory works led by the ACPR suggest adopting a dual approach:

- The first facet is analytical. It combines analysis of the source code and of the data with methods (if possible based on standards) for documenting AI algorithms, predictive models and datasets.
- The second facet is empirical. It leverages methods providing explanations for an individual decision or for the overall algorithm’s behaviour, and also relies on two techniques for testing an algorithm as a black box: challenger models (to compare against the model under test) and benchmarking datasets, both curated by the auditor.

Such a multi-faceted approach is suitable both for internal auditors and for a supervisory authority, however the latter faces specific challenges due to the scope of its mission. In order to effectively audit AI systems, it will need to build both theoretical and hands-on expertise in data science, while developing a toolkit for the specific purpose of AI supervision.

Public consultation

The analysis presented in this document is subject to public consultation. The objective is to submit to financial actors and other concerned parties (researchers, service and solution providers, control authorities, etc.) guidelines sketched herein for feedback, and more broadly to gather any useful comment, including on supervisory authorities’ best practices.

2. Introduction

2.1. Methodology

Following initial work and a December 2018 public consultation on the role of Artificial Intelligence (AI) in finance, in 2019 the ACPR's Fintech-Innovation Hub undertook exploratory works with a small number of voluntary financial institutions in order to shed light on the issues of explainability and governance of AI in the sector. These exploratory works resulted in the avenues for reflection presented in this document. The technological spectrum considered here is detailed in appendix 7.

Financial actors are, as evidenced by the APCR's first public consultation, particularly eager for regulatory guidance pertaining to AI[1]. Indeed, this technology generates opportunities as well as risks – operational and otherwise. One of a supervisory authority's tasks is to provide such guidance, along with practical implementation guidelines, with the aim of balancing freedom of innovation with regulatory compliance and responsible risk management.

2.2. Exploratory works

The main goal of these exploratory works was to produce lines of thought on three topics, each related to the ACPR's main missions and detailed in what follows.

In each topic, the Fintech-Innovation Hub conducted a deep-dive analysis with voluntary actors in a two-fold way:

- In each case, meetings to present the AI algorithms in question along with the main explainability and governance challenges.
- In the case of the "primary" workshops, a more technical phase involving data scientists on both sides, exchanging on relevant methods and tools, including review sessions of the source code and ML models developed by the actor.

These workshops are briefly described hereafter. Appendices provide more detailed, anonymized descriptions.

2.2.1. Topic 1: Anti-money laundering and combating the financing of terrorism (AML-CFT)

This topic's key issue was whether AI can improve financial transaction monitoring, either by complementing or by replacing traditional threshold mechanisms and other business rules. To tackle this challenge, workshop participants introduced ML algorithms able to generate alerts (in addition to traditional, rule-based systems already in place): those alerts are directly sent for review to "level 2" teams[2], which streamlines and secures the alert review workflow. The resulting operational gain is proven while the governance of two key AML-CFT processes studied in these workshops (namely declarations of suspicion and freezing of assets) needs to be re-examined in light of the human intervention required by those processes, and of the continuous monitoring required for ML algorithms.

[1] See also Cambridge Judge Business School, 2020.

[2] Level 2 is in charge of compliance: see section 5.1.2 for explanations on the typical organization of internal controls.

2.2.2. Topic 2: Internal models in banking and insurance

This topic’s key issue was to determine how –and under which conditions – AI can be used in the design of internal models.

Rather than studying internal models as a whole, the workshops have focused on credit granting models; both are related insofar as scores produced by those models can also be used to build risk classes, from which RWAs (Risk-Weighted Assets) are computed.

The workshops involved two actors: a banking group which designs and implements its credit scoring models internally, and a consulting firm which provides a development platform for hybrid (ML- and rule-based) models, tested in this case on the computation of the probability of default. Both application scenarios demonstrated how introducing ML impacts governance: the initial validation process becomes more technically-oriented, monitoring tools become a requirement for internal review, and explanatory methods have to be integrated into continuous oversight as well as into audit processes.

2.2.3. Topic 3: Customer protection

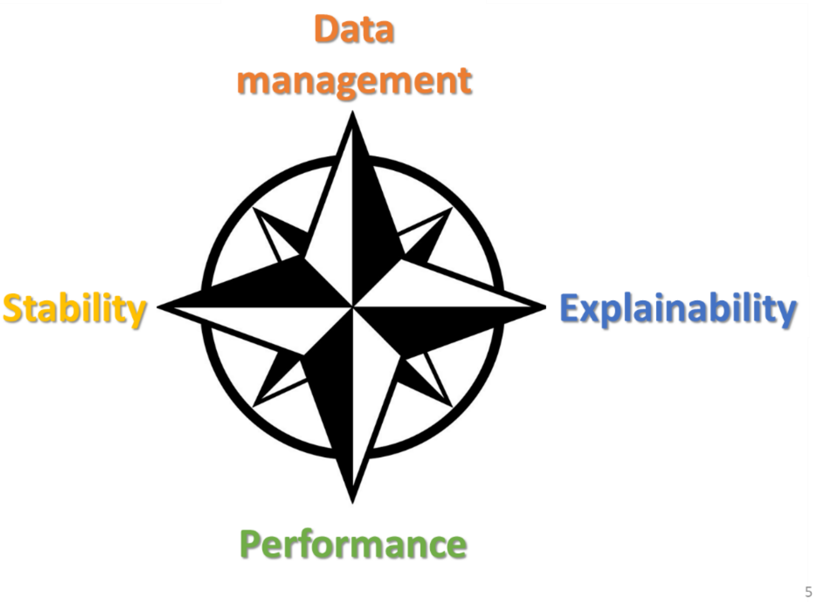
This topic’s key issue was to ensure that AI algorithms, when used in the context of non-life insurance product sales, always take into account the client’s best interests.

The ML model studied on this topic aimed at producing prefilled quotes for home insurance products. That use case involved two main compliance requirements: fulfilling the duty to advise so as to properly inform the customer, and offering a non-life insurance product which is consistent with the requirements and needs expressed by the customer.

3. AI design and development principles

We suggest four evaluation principles for AI algorithms and models: appropriate data management, performance, stability, and explainability.

The objectives represented by these principles need to be mutually balanced: simultaneously maximizing all of them is usually impossible. They can be viewed as the cardinal points of a compass by which to guide the design and development of an AI algorithm:



3.1. Appropriate data management principle

Evaluating the compliance of an algorithm and of its implementation requires covering a large spectrum of requirements. At the very core of those compliance requirements lies the proper management of data during each stage of the design and implementation process of an AI algorithm, as described in this section.

Input data

Defined as data fed to an algorithm during its conception, input data comprises reference data, training data, and test (or evaluation) data. Proper management of input data is sometimes governed by sector-specific regulation, for example data completeness and data quality requirements in the banking sector are prescribed by prudential norm BCBS 239[3].

Data governance (Dai, 2016) is an essential function within any financial organization with a number of data-driven business processes. Setting up a proper data governance for an AI algorithm will not work if data sources fed to it are inappropriately managed, for example if they are fragmentary, anecdotal, insufficiently durable, can be tampered with, or if the organization does not control their lifecycle.

[3] Basel Committee on Banking Supervision's standard number 239 is an international standard, whose subject title is *"Principles for effective risk data aggregation and risk reporting"*.

Pre-processing

Evaluation of an ML-based system also needs to take into account operations performed on input data prior to the machine learning phase itself. Pre-processing may have an impact on the resulting model's performance (for example by over- or under-sampling training data) as well as on its ethical acceptability (for example by excluding protected variables from training data).

Post-processing

Finally, evaluation should also include operations performed on the output (i.e. predictions or decisions) of the model produced by the ML algorithm. Such post-processing may have a significant impact as well, such as in the case of methods aiming to remove or reduce discriminatory biases[4] from already trained models – for example by cancelling out the dependency of predictions made by a probabilistic model on sensitive variables (Kamishima, 2012)[5].

3.1.1. Regulatory compliance

Regulatory compliance often includes requirements of two kinds:

- Compliance with regulation pertaining to data protection and individual privacy, starting with GDPR in Europe.
- Taking into account regulatory requirements specific to a domain or use case. For example in insurance, it is prohibited to steer the sales process based on the customer's capacity to pay: the offer needs to be at least consistent with the demands and needs expressed by the customer – not driven by the maximization of sales revenue from insurance products.

Compliance with the first category of requirements can be assessed by well-proven methods: undesired biases can be detected, prevented or suppressed (by operating at any of the aforementioned stages: input data, pre- or post-processing), dependency on sensitive variables (whether explicitly or implicitly present in the data) can be suppressed, etc.

The second category of compliance requirements, those which are sector-specific, often goes beyond the scope of data management: this is the case of the obligation of means and of the performance obligation in AML-CFT, which call for suitable explanatory methods.

Another example will illustrate the stakes of sector-specific regulation in further detail: that of an ML system put in production in an insurance organization, which aims to target high-priority prospective customers for a marketing campaign regarding a multi-risk insurance contract. The IDD (Insurance Distribution Directive), a 2016 European directive, introduced principles close to the equity principle described in the next section: insurance product distributors should "always act honestly, fairly and professionally in accordance with the best interests of their customers." Therefore, ML is only allowed for customer targeting on the condition that the criteria used are based on the needs fulfilled by the

[4] The polysemy of the term "bias" should be noted. It sometimes refers to a statistical, objective characteristic of a predictor or estimator, other times to an unfair or unequal treatment whose polarity and importance are subjective and of an ethical or social nature. The presence of a statistical bias may lead to a fairness bias, but this is neither a sufficient nor a necessary condition.

[5] The issue of discriminatory biases is not specific to AI either. The risk exists in any statistical model, e.g. it is documented in the literature on "redlining" in banking economy. However that risk is amplified by the use of an ML algorithm, in addition some detection and mitigation techniques for that risk are also ML-specific.

product – and not on the customer's capacity to subscribe[6]. The challenge is thus to correctly appreciate the prospective customers' insurance needs, which are much more difficult to evaluate for an algorithm than for a human. When ML is used, this requires using larger datasets (in breadth and in depth), which in turn generates or increases data-induced risks such as implicit correlations with the capacity to subscribe (which are difficult to detect) and more generally undesired biases (themselves often latent, see next section). In short, using ML to implement a customer targeting system for marketing should be conditioned upon mastering those risks and deploying tools to detect and mitigate them.

3.1.2. Ethics and fairness

Besides constraints stemming from sector-specific and cross-cutting regulations, ethical issues lie at the core of the ever-increasing usage of AI in business processes which impact individuals and groups of people. Those issues include social and ethical concerns in the broadest sense, and particularly questions of fairness raised by any automated or computer-aided decision process.

Ethics guidelines published by the European Commission (European Commission High-Level Expert Group on AI, 2019) illustrate both the importance of ethical issues and the blurred boundaries they share with the other principles described in this section:

1. Human agency and oversight
2. Technical Robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability.

These guidelines underline the broad spectrum of challenges related to ethics and fairness in AI. Specifically, the analysis of biases – especially those of a discriminatory nature – is an active research domain, which schematically comprises the following stages:

- Carefully defining what constitutes a problematic bias – whether a classification bias or prediction bias, or an undesired statistical bias already present in input data – and the metrics enabling to characterize and quantify such biases, including via explanatory methods (Kamishima, 2012) ;
- Determining to what extent biases present in the data are reflected, if not reinforced, by AI algorithms
- Lastly, mitigating those biases whenever possible, either at the data level or at the algorithm level.

Exploratory works conducted by the ACPR, along with a broader analysis of the financial sector, showed that bias detection and mitigation were at an early stage in the industry: as of now, the emphasis is put on internal validation of AI systems and on their regulatory compliance, without pushing the analysis of algorithmic fairness further than was the case with traditional methods – in particular, the risk of reinforcing pre-existing biases tends to be neglected. This blind spot, however, only reflects the relative lack of maturity of AI in the industry, which has been introduced primarily

[6] Indeed, the conception of an insurance product should start by defining a target market, based on characteristics of the group of customers whose needs are fulfilled by the product.

into the less-critical business processes (and those which bear little ethics and fairness risks): it can thus be anticipated that the progressive industrialization of additional AI use cases in the sector will benefit the currently very active research on those topics.

APPROPRIATE DATA MANAGEMENT

All data processing should be as thoroughly documented as the other design stages of an AI algorithm (source code, performance of the resulting model, etc.) This documentation enables risk assessment in the areas of regulatory compliance and ethics, and the implementation of tools for detecting and mitigating undesired biases, if need be.

3.2. Performance principle

Performance of an ML algorithm can typically be assessed using two types of metrics:

- Predictive performance metrics. For instance AUC (Area Under the Curve) – or alternatively F1 score – can be applied to algorithms which predict credit default risk of a physical or moral person. Such metrics can be categorized as KRI or *Key Risk Indicators*.
- Business performance metrics, which can be categorized as KPI or *Key Performance Indicators*. Two points of attention for such metrics are their consistency with the algorithm's objectives and their compatibility with its compliance requirements[7].

Importantly, algorithmic performance of an AI algorithm is not a standalone objective: it needs in particular to be weighed against the explainability principle. Subsequent sections will show that the adequate explanation level depends, for a given scenario, on multiple factors and on the explanation's recipients. The choice of an explanation level in turn induces constraints on technological choices, notably on the "simplicity" of the selected algorithm.

An overview of the fundamental trade-off driving such choices is given in appendix 10.1.1.

ALGORITHMIC PERFORMANCE

Performance metrics of an ML algorithm have to be carefully selected, so as to evaluate the technical efficacy of the algorithm or alternatively its business objectives. The inherent trade-off between the algorithm's simplicity and its efficacy has to be taken into account.

3.3. Stability principle

The stability principle consists of ensuring that an ML algorithm's performance and its main operational characteristics are consistent over time. Expectations in terms of stability are all the more important

[7] It is for instance highly likely that maximizing sales revenue from insurance products is an inappropriate metric for an ML algorithm used as part of the sales process: it might indeed introduce into the algorithmic process the kind of conflicts of interest which regulation precisely aims at preventing.

in the case of ML, as the dimension of data (in order to make up predictor variables) tends to be much larger than in traditional predictive or decisional algorithms[8].

Three major sources of instability are herein identified. As of now, AI algorithms in production in the financial sector seldom take into account these instability sources, neither individually nor for their overall effect. This may be due to the relative lack of maturity of AI engineering and operational processes, and thus subject to change in the future. However ML instability risks should not be neglected since they generate significant operational and compliance risks. Hence a few mitigation methods are suggested in the following for each of them.

3.3.1. Temporal drift

Firstly, stability of an ML model implies absence of drift over time. This is essential since the distribution of input data might deviate sufficiently to degrade the model's performance as well as other characteristics (such as compliance-related aspects or the absence of bias) – especially if it is not periodically re-trained.

This temporal drift can be detected using somewhat classical monitoring and alert generation mechanisms, which should however be built upon appropriate drift indicators and a well-tried infrastructure. A key point in this regard is that temporal drift of a model is often linked to the evolution of the training database, hence the very first stage when designing a drift monitoring tool – before even taking into account data processing – consists of detecting structural changes in those input data.

3.3.2. Generalization

Stability of an ML model can also be understood as robustness, in the sense of generalization power when confronted with new data[9]. A lack of generalization power may have gone undetected during the model validation, for example because test and validation datasets – though dissociated from training data as they may be ("out-of-time" testing and "out-of-distribution" testing) – almost inevitably differ from the real-world data fed to the model in production.

Lack of generalization power may be detected and (at least partially) remedied during the model design and parameterization stages. However the resulting model should be subjected to continuous monitoring because, just like temporal drift is ultimately inevitable, the model's performance can never be guaranteed to generalize sufficiently well to previously unseen data.

3.3.3. Re-training

Lastly, re-training an ML model, whether periodically or on a quasi-continuous basis, does not solve all instability issues, since it results at the very least in non-reproducible decisions on a given input

[8] It is even one of Big Data's main characteristics, and a situation where techniques such as neural networks particularly shine. Generally speaking, the predictive power of a classification model can be shown to increase with the number of variables up to a certain point, after which it degrades – a phenomenon called Hughes' peak (Koutroumbas, 2008) and associated to the "curse of dimensionality". Dimensionality reduction is actually a very common concern in ML (Shaw, 2009).

[9] Generalization power and predictive bias are the two key criteria to balance when designing and tuning a predictive model. Generalization is inversely proportional to the model's variance, hence this arbitrage is referred to as bias-variance trade-off: low bias is usually associated to high performance on training and test data, whereas low variance implies that the model generalizes well to new data.

between subsequent versions of the model. The main consequence of this instability source over the course of the model's lifecycle (thus punctuated by re-training phases) is a lack of determinism in the overall system. This can become a problem when a particular decision must be reproduced (for example to comply with GDPR's right to access and opposition), possibly accompanied by an explanation (which can be produced by one of the explanatory methods described hereafter).

This instability source, when it cannot be mitigated via a low-enough re-training frequency, can at least be compensated by properly archiving all subsequent versions of an ML model used in production.

STABILITY

Potential instability sources which may affect AI algorithms deployed in the organization over time should be identified. For each such source, associated risks (operational, compliance risks, or otherwise) should be assessed, and proportionate detection and mitigation methods implemented.

3.4. Explainability principle

Of the four principles exposed here, explainability is the one most distinctive of AI systems compared to traditional business processes.

3.4.1. Terminology

Notions of algorithmic explainability, transparency, interpretability, and auditability are closely related:

- Transparency is but a means (albeit the most radical) to make decisions intelligible: it implies access to an ML algorithm's source code and to the resulting models. In the extreme case of complete opacity, the algorithm is said to operate as a black box.
- Auditability means the practical feasibility of an analytical and empirical evaluation of the algorithm, and aims more broadly at collecting explanations on its predictions, as well as evaluating it according to the aforementioned criteria (data management, performance, and stability).
- The distinction between explainability and interpretability has been the subject of numerous debates, which are summarized in appendix 9. The term "explainability" is often related to a technical, objective understanding of an algorithm's behaviour (and would thus be more suitable for auditing), whereas interpretability seems more closely associated with a less technical discourse (and would thus primarily target consumers and other individuals impacted by the algorithm).

3.4.2. Objectives

Explanations pertaining to an AI algorithm generally address the following questions:

- ➢ What are the causes of a given decision or prediction?
- ➢ What inherent uncertainty does the model carry?
- ➢ Are the errors made by the algorithm similar to those due to human judgment?
- ➢ Beyond the model's prediction, what other pieces of information are useful (for example to assist a human operator in making the final call)?

Thus the objectives of an explanation vary greatly, especially depending on the stakeholders considered:

- Providing insights to domain experts and compliance teams.
- Facilitating the model's review by the engineering and validation teams.
- Securing confidence from the individuals impacted by the model's predictions or decisions.

An overview of the fundamental trade-off driving the technical choice of an algorithm based on the types of explanations required is given in appendix 10.1.2.

3.4.3. Properties

An ideal explanation should have the following properties:

- **Accurate**: it should describe as precisely as possible the case studied (for a local explanation) or the algorithm's behaviour (for a global explanation).
- **Comprehensive**: it should cover the entirety of motives and characteristics of the considered decision(s) or prediction(s).
- **Comprehensible**: it should not require excessive effort in order to be correctly understood by its intended recipients.
- **Concise**: it should be succinct enough to be grasped in a reasonable amount of time, in accordance with the time and productivity constraints of the encompassing process.
- **Actionable**: it should enable one or more actions by a human operator, such as overriding a prediction or decision.
- **Robust**: it should remain valid and useful even when input data are ever-changing and noisy.
- **Reusable**: it should be customizable according its intended recipients.

In practice, not all of these qualities are simultaneously achievable. Besides, as previously mentioned, they have to be balanced against other principles – notably performance. Thus these properties should rather serve as comparison criteria between explanations provided by various methods, so as to select the one most appropriate to a specific use case.

3.4.4. Explanation levels

For simplicity's sake, we adopt hereafter the term "explainability" rather than "interpretability" to describe the broader concept (cf. section 9 on the terminology). Algorithmic explainability aims to demonstrate:

- On the one hand, *how* the algorithm operates (which roughly matches the common meaning of algorithmic transparency).
- On the other hand, *why* the algorithm makes such and such decision (in other words an interpretation of those decisions).

A key challenge of the "why" question is the auditability of an ML algorithm. As for the "how" of explainability, associated challenges include:

- For human operators who interact with the AI system: to understand its behaviour
- For individuals affected by the system's predictions or decisions (such as customers in a sales context): to understand the underlying motives
- For those who designed the system or are tasked with checking its compliance: to assess its social and ethical acceptability, in order (among other things) to prove the absence of discriminatory bias in its decision-making process.

The concept of an explanation level introduced here attempts to summarize in a single metric the depth of an explanation[10]. This metric exists on a continuum, along which we define a four-level scale of qualitatively distinct levels, which are described in the following.

Level-1 explanation: observation

Such an explanation answers technically-speaking the question *"How does the algorithm work?"* and functionally-speaking the question *"What is the algorithm's purpose?"* This level can be achieved:

- Empirically, by observing the algorithm's output (individually or as a whole) as a function of input data and of the environment
- Analytically, via an information sheet for the algorithm (cf. appendix 11.1), the model, and the data used, without requiring the analysis of the code and data themselves.

Level-2 explanation: justification

Such an explanation answers the question: *"Why does the algorithm produce such a result?"* (in general or in a specific situation). This level can be achieved:

- Either by presenting in a simplified form some explanatory elements from higher levels (3 and 4), possibly accompanied with counterfactual explanations (cf. appendix 11.3).
- Or by having the ML model itself it has been trained to produce (cf. appendix 11.2).

Level-3 explanation: approximation

Such an explanation provides an – often inductive – answer to the question *"How does the algorithm work?"* This level of explanation can be achieved, in addition to level-1 and 2 explanations:

- By using explanatory methods which operate on the model being analysed (cf. appendix 11.3).
- Via a structural analysis of the algorithm, the resulting model and the data used. This analysis will be all the more fruitful if the algorithm is designed by composition of multiple ML building blocks (hyper-parameter tuning or auto-tuning, ensemble methods, boosting, etc.).

[10] This concept is thus by definition an over-simplification of the quality of an explanation. It aims at facilitating the choice of a target explainability level, without eliminating the need for a multi-dimensional analysis of the explanations provided.

Level-4 explanation: replication

Such an explanation provides a demonstrable answer to the question *"How to prove that the algorithm works correctly?"*

This level of explanation can be achieved, in addition to level-1 to 3 methods, by detailed analysis of the algorithm, model and data. In practice, this is feasible only by doing a line-by-line review of the source code, a comprehensive analysis of all datasets used, and an examination of the model and its parameters.

It should be noted that each level characterizes an explanation (or a type of explanation), rather than an ML algorithm or model. Strictly speaking, it is about *the level of intelligibility of the explanations provided by the system*, not about the intrinsic explainability of the system. Thus, a highly explainable model such as a decision tree might lend itself to a level-4 explanation (by thoroughly detailing all its branches), but also to a level-1 explanation (by simply stating that it is a decision-tree predictor operating on a given set of input variables). The latter would suffice in a case where the fine-grained behaviour of the model does not need to be – or *must* not be – disclosed.

Under a more technical perspective, appendix 10.2 examines in further detail the technical feasibility of higher-level (3 or 4) explanations: it presents an important hurdle to overcome (software dependencies) along with a path to reach level 4 (replication).

The next sections describe two factors – among a number of them – driving the explanation level required from an AI algorithm, especially in the financial sector: on the one hand the intended recipients of the explanation, on the other hand the risk (both its nature and its severity) associated to the considered process. Thus, the same algorithm might require a higher explanation level when its inner behaviour also needs to be captured and/or when the explanation is provided in a particularly sensitive context.

3.4.5. Recipients of the explanation

The first key influence factor of the expected explanation level is the type of recipient targeted. This is because the relevant form under which an explanation should be proposed in order to be effective depends both on their technical and business proficiency and on their intrinsic motives for demanding an explanation.

Hence different explanation levels could be applied to the same algorithm depending on whether the explanations serve an end user (who tries to check that they have not been treated unfairly by the system, and for whom an explanation has to be intuitively intelligible) or an auditor (who needs to understand the system's technical architecture in detail and who is subjected to rigorous regulatory requirements).

We hereafter describe three kinds of recipients for an explanation, and suggest each time what an appropriate form of explanation looks like.

Customer or consumer: simple explanations

An example of explanation intended for a customer occurs in the context of insurance product sales: the duty to inform makes it mandatory to explain to prospective customers why they were offered a given insurance product and not another one, furthermore those motives need to be cantered around the consistency of the contract (in the case of non-life insurance) or its adequacy (in the case of life insurance).

The nature and terms of this explanation must therefore be intelligible and satisfactory with regard to the consumer (who cannot be required to master the intricacies of the sales process, nor the implementation of the underlying algorithm).

Continuous monitoring: functional explanations

Internal review teams, particularly those tasked with continuous monitoring, need to verify the model's efficacy with respect to business objectives.

The focus in this case is put on the performance of the process involving AI, rather than on its internal mechanics, thus the explanation given should be of a functional nature.

Auditor: technical explanations

Thirdly, an auditor must ensure that the algorithm's implementation is consistent with respect to its specifications, including in terms of regulatory compliance and of technical requirements.

This entails, for example, verifying how an ML model is produced, but also checking the absence of discriminatory bias in that model. Therefore the explanation given must be technically accurate and as representative as possible of the audited model.

3.4.6. Associated risks

The second factor of influence on the required explanation level is the risk associated to the (total or partial) replacement of a human-controlled process by an AI component.

The nature and severity of that risk are highly variable, as shown by the following examples:

- **AML-CFT**: a process such as freezing of assets, which is subjected to a performance obligation, bears an increased level of risk when AI is introduced, not only by dint of its critical role, but also because its evaluation then depends on comparing human and algorithmic efficiency. More precisely still, the risk will be particularly elevated in a continuous monitoring or audit situation (which has to assess that relative efficiency) and more moderate for a daily user of the algorithm who keeps performing the same controls as when using a traditional transaction monitoring system.
- **Internal models**: introducing ML into e.g. the computation of solvency margins of a banking institution has a direct impact on the assessment of its solvability risk, therefore the team who designs the institution's internal model will expect a satisfactory level of explanation for the results of those computations.
- **Insurance**: the insurance contract sales process has its own regulation, which imposes among other things a duty to inform and the personalized presentation of reasons and justifications to the customer, if need be. Conversely, *ex-ante* customer segmentation in the insurance sector relies mainly on accuracy objectives, without the same requirement in terms of explainability.

EXPLAINABILITY

For each use case of AI, the impacted business processes should be determined and the roles filled by the AI component should be detailed. The types of recipients targeted by an explanation can then be described, along with the nature of the associated risks. That entire context dictates the level and form of an appropriate explanation for the AI algorithm, which must be agreed upon by all stakeholders in the algorithm's governance.

3.4.1. Examples of explanation levels by use case

We attempt here to illustrate those somewhat abstract definitions of explanation levels and of their driving factors through a few concrete use cases – all of which have been deployed by financial entities, and some of which were analysed during the exploratory workshops conducted by the ACPR.

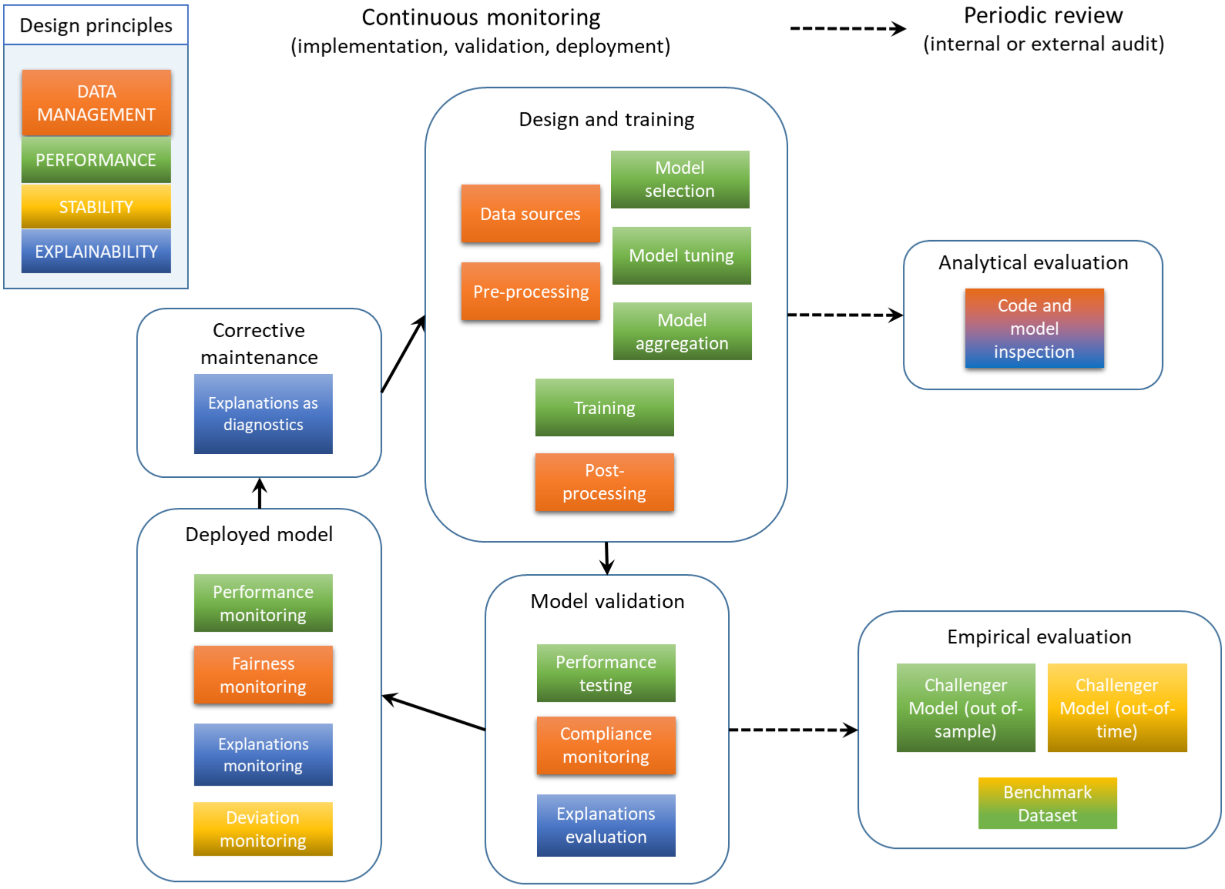
For each use case, the following table suggests an explanation level as a function of the aforementioned criteria (targeted recipients and associated risk). Those suggestions are based on our initial market analysis, whose observations the present document aims to validate or correct (see consultation in section 6).

Use case			Explainability criteria			Appropriate explanation level
Domain	Busines process	AI functionality	Explanation recipients	Context	Associated risk	
Insurance contracts	Contract management	Compensation proposal	Customer	Compensation process	Operational risk (customer satisfaction)	1
			Internal control	Daily oversight of the process	- Operational risk - Compliance risk (contract honoring) - Financial risk	2
			Auditor	Evaluation of the algorithm	- Operational risk - Compliance risk (contract honoring) - Financial risk	3
	Sales proposal	Quote prefilling	Customer	Online quote request	Compliance risk (customer misinformation, failure to perform duty to inform, discriminatory biases…)	2
			Internal control or internal auditor	Compliance checking	Compliance risk (customer misinformation, failure to perform duty to inform, discriminatory biases…)	3
Internal models	Model design	Computation of solvency margins	Validation team	Model and model update policy validation	- Solvency model risk - Compliance risk	4
			Administrative, management, and supervisory bodies	Model approval	- Solvency model risk - Compliance risk	2
AML-CFT	Freezing of assets	Alerting	Level-2 agent	Alert analysis	None (if the analyst's methodology is not modified by the algorithm)	1
			Internal control	Continuous monitoring	- Operational risk (false negatives / false positives) - Compliance risk (performance obligation)	2
			Auditor	Periodic inspection	- Operational risk (false negatives / false positives) - Compliance risk (performance obligation)	3

4. Evaluation of AI algorithms

The following diagram represents the lifecycle of an AI algorithm and of the resulting model, from the design and training phases to its use in production – and possible iterations to the learning stage, for instance upon patching the algorithm. It attempts to put in perspective those implementation stages with the appropriate validation steps, whether continuous or periodic, internal or external.

It also aims to show how each stage in the lifecycle benefits from a suitable evaluation process, based on the four principles previously mentioned, namely data management, performance, stability, and explainability. Finally, it illustrates the multifaceted approach to evaluation detailed in section 5.5.1, which combines analytical and empirical evaluation.



EVALUATION OF AI

The lifecycle of AI algorithms introduced into each business process should be detailed so as to list, at each stage, which design and development principles (data management, performance, stability, explainability) apply in particular, and which evaluation method is appropriate for that stage.

5. Governance of AI algorithms

Introducing ML algorithms into the financial sector often aims, be it via descriptive or predictive methods, to automate or improve (e.g. by customizing it) a decision-making process which used to be performed by humans. Therefore, the governance of those algorithms requires careful consideration of the validation of each of these decision-making processes. In particular, their regulatory compliance objectives as well as their performance objectives are only achievable through a certain level of explainability and traceability.

The following governance concerns need to be taken into account as early as the design phase of an algorithm[11]: integration of AI into traditional business processes; impact of this integration on internal controls, specifically on the role assigned to humans in the new processes; relevance of outsourcing (partially or fully) the design or maintenance phases; and lastly, the internal and external audit functions.

5.1. Governance principles in the financial sector

General governance principles applicable to any financial institution include the description of the “control culture” policy implemented in the organization, the presentation of ethical and professional norms it promotes, along with the steps taken to guarantee proper implementation of those norms and the process in case of failure to do so. In addition to those principles, other procedures should be documented, such as how to detect and prevent conflicts of interest.

In this context, the most relevant elements of governance when introducing AI into business processes appear to be the operational procedures within those processes, the extension of segregation of duties to the management of AI algorithms, and the management of risks associated to AI. These elements are briefly described in this section.

5.1.1. Operational procedures

Operational procedures should be adjusted to the different activities performed, communicated, and periodically updated, for example via a clear written documentation. Their main goals are to describe how the various levels of responsibility are assigned, the resources devoted to internal control mechanisms, the risk measurement, mitigation and monitoring systems implemented, and the organization of compliance monitoring. Those procedures also list rules relative to IT security and to business continuity planning.

5.1.2. Segregation of duties

There are no organizational standards relative to internal controls and risk management, only methods which have been tried and tested when implementing such functions (COSO, Cobit, Risk IT, etc.).

[11] This document does not address another governance issue, which should nevertheless precede any decision to adopt a technical tool – independently of its usage of AI and of its business application –, namely the cost/benefit analysis. In other words, only governance questions specific to the usage of AI in the financial sector are considered herein.

Nevertheless, internal control mechanisms conventionally comprise multiple levels of control, so as to follow the “four-eyes principle”. Classically those levels are[12]:

- A level-1 control, within the business units which conduct their activities or perform their duties in a controlled manner.
- A level-2 control, exercised by the unit managers or directors, or in more complex organizations by teams responsible for internal controls (also referred to as internal oversight).
- A level-3 control, exercised by the internal audit directorate, which aims to guarantee the proper implementation of control mechanisms by periodically reviewing their operational accuracy.

A clear segregation of duties must exist between business units which commit operations, and those which record and monitor operations on an ongoing basis.

5.1.3. Risk recognition and assessment

Organizations should perform a risk mapping, which must be periodically updated and evaluated, so as to develop a coherent and comprehensive view of risks. They should also define and regularly promote a solid, consistent risk culture dealing with risk awareness and with risk-taking behaviour. Lastly, they should implement systems and procedures to guarantee a cross-cutting, prospective risk analysis.

5.2. Integration in business processes

One of the main challenges for the governance of AI algorithms is their integration in existing processes. Key factors to take into account are the role played by the algorithms within a business process, the engineering methodology used, and who the end users are.

5.2.1. Role of AI

The roles played by AI components in business processes are highly variable.

The primary AML-CFT workshop (section 8.1) illustrates how the function of an ML model can be operational, even business-critical: in that case, its role is to route certain alerts triggered by financial transactions with a particularly high estimated risk directly toward level 2 (Compliance), thus inducing an operational risk in case the Compliance team becomes overloaded. The critical function of the AI component is also elevated in this case by the constraint of real-time operation: suspicious transactions should be detected and reported with as small a lag as feasible.

Conversely, the incorporation of ML into the prospective customer selection process for the purposes of commercial canvassing or cross-selling is not truly disruptive, and does not incur any significant change in the business process.

5.2.2. AI engineering methodology

The definition of an appropriate engineering process for ML varies greatly depending on the business process and on how the models are used. Two examples shall illustrate this variety of situations:

[12] The 2013 “CRD-IV” European directive (Capital Requirements Directive) defines the basis for such an organization within financial institutions.

- When ML is used by marketing teams (a common case, although not covered by the exploratory works described herein), significant room for manoeuvre is granted to model building, which is often an iterative process since one-off model deployment is used e.g. to feed a marketing campaign.
- Conversely, ML use cases studied in the ACPR's workshops require a more systematic engineering process, closer to the best practices adopted by the software industry: build automation, reproducibility of releases, QA (quality assurance) process, monitoring of the models deployed in production (including their stability over time). The engineering process should thus meet more stringent requirements in this latter case.

Thus the AI engineering process can vary from a one-off build-and-deploy mode, through an iterative build process, all the way to a continuous process which can also be fully automated, typically using CI/CD[13].

As for the delivery mode of the AI system itself, it is also variable between a manual delivery process where only final artefacts (i.e. the ML models) are retained to be put in production, and at the other extreme delivering the entire datasets and intermediate results from the algorithm's execution and model-building stages. A middle ground between those two approaches is the "managed services" approach offered by the consulting firm which participated in the workshop on probability of default (section 8.4), which is composed of two elements: on the one hand a model engineering workbench which follows a systematic (albeit not fully automated) model-building approach but is controlled by the solution provider, on the other hand an information sharing platform which enables the customer who uses the ML model to perform a complete review of the engineering process, and provides an audit track independent from the execution of that process.

AI ENGINEERING METHODOLOGY

The AI engineering process should be designed to cover the entire algorithm lifecycle. Depending on the use case, a systematic approach may be necessary, in accordance with principles of model-building automation, build reproducibility, quality assurance, and monitoring of the engineering workflow.

In any case, full traceability of the AI design and engineering process should be guaranteed.

5.2.3. AI end users

The impact of the introduction of AI in a business process primarily depends on who its end users are – as opposed to personnel responsible for its internal control whose role will be examined in the next section. Those end users may be internal such as marketing teams and business unit managers, or external such as clients and prospects.

[13] CI/CD (Continuous Integration / Continuous Deployment) refers to general software engineering principles based on automating the entire design and development process, which enables more frequent product releases than traditional methods allow, without trading off their quality. This methodology is closely related to agile methods as well as the DevOps approach, which associates the roles of software development and IT operations.

In particular, maintaining the quality expected from the process requires examining whether a particular form of explanation should be provided to end users so as to clarify and motivate the decisions and predictions impacting them.

Types of end users

In the case of integrating an ML component into an AML-CFT workflow (see details on these works in the appendix), end users are level-1 and level-2 teams:

- Verifications performed by the Compliance team need to be adequate to this new approach (which requires mastering the underlying technology).
- Model validation needs to be performed much more frequently than e.g. in the case of capital requirements models, since drifts may occur in real time here (for example a false positive rate which deviates from the norm), hence monitoring of the model must itself be feasible in (quasi) real time.

In the case of the workshop on customer protection (section 8.5), prefilled quotes for a home insurance contract are delivered to the customers themselves, which requires explaining the reasons for offering a specific product. These reasons must be in line with the customers' requirements and needs.

Human-machine interactions

It is essential that end users of an algorithm, insofar as they are internal users tasked with ensuring the accuracy and quality of a business process, remain independent from the machine. This is because human experts are able to spot manifest errors made by the algorithm, which also offers the benefit of contributing to its performance and stability (i.e. two out the four design principles presented in this document).

AI also provides additional leverage to check the absence of systematic biases or temporal drift in the decisions made – or the advice given – by an automated process in finance: in such a situation, introducing ML into processes enables to decrease the operational risk.

Human intervention in a decision-making process delegated to software is not inconsequential, as it introduces a new kind of risk: the downside of enabling a human operator to validate the decisions is that they may become liable, especially in cases where they contradict the algorithmic result rather than confirm it. Besides, humans sometimes modify their own behaviour when interacting with a machine: they may tend to systematically follow the algorithmic results, including the erroneous ones, rather than engaging their liability by rebutting them.

This issue of independence from the algorithm and responsibility towards its decisions are of course related to the explainability principle, since a human operator needs to understand the mainspring of a given decision in order to, if need be, counter it with a more appropriate one.

Lastly, human intervention might introduce bias – desired or not – into the explanations associated to an algorithm's output, whenever the explanations provided or amended by the operator become disconnected from the actual underlying factors which led to that output: the explanations become distorted or overridden, furthermore transparency is no longer guaranteed for the algorithm, which may hide some of its deficiencies. A basic recommendation appears relevant in this regard, namely not allowing human intervention to define or formulate the algorithm's explanations.

AI END USERS

The scope and conditions of human intervention in AI-driven business processes should be well-defined. In particular, AI integration into those processes should be planned according to end users' needs. If end users include both internal and external individuals, the respective forms of algorithmic explanations appropriate to each of those should be articulated.

Algorithmic results may also have to be submitted to a human validation process. This validation should be governed by rules documented as part of the internal control procedures, both because human responsibility becomes engaged and because the algorithm may modify human behavior and judgment.

5.3. Internal control system

The other major impact of introducing AI algorithms pertains to the continuous validation of those algorithms, and specifically internal control procedures.

5.3.1. Organization of internal controls

Monitoring algorithmic performance and detecting its potential drift over time requires a different design of the human validation process. For example, the AML-CFT workshop (section 8.1) illustrated how the partial replacement of level-1 operators by an ML algorithm may decrease the capacity to evaluate the process efficiency in the future, at least in terms of false negatives (alerts not raised by the system, and thus corresponding to transactions which will not be analysed by the human eye): this is why some of those operators have been assigned to manual labelling in parallel with the algorithm's execution, thus continuously yielding new training data.

As for the organizational structure of internal controls, an ML algorithm often aims to replace (partially or fully) the tasks performed by the level-1 team (reviewed by their hierarchy) and/or level-2 team (in charge of compliance checks), but probably not level-3 (in charge of internal controls) – although nothing precludes this automation stage to be achieved in a more distant future. The users of the algorithm's output are thus – as should be – not part of the team tasked with monitoring its behaviour, nor are they its designers.

ORGANIZATION OF INTERNAL CONTROLS AND AI

Internal control procedures of AI algorithms should, to the extent possible, involve both technical specialists and domain experts. Indeed, monitoring those algorithms requires initial technical validation of the components involved, their continuous monitoring, and adequate management of compliance risks generated or reinforced by the ML.

5.3.2. Initial functional validation

In the case of the workshop on credit scoring models (section 8.3), a pre-deployment model validation process was defined, with the involvement of technical teams (in charge of building and validating models, both locally and globally within the banking group in question) and of the compliance and risk management department.

In particular, any model deployment (be it a model creation or the patch for an issue affecting an already-deployed model) requires validation by the group-level Risk Committee, among other things to approve the chosen risk management strategy. The use of ML within these models is thus considered and evaluated by stakeholders across the organization: by technical experts, domain specialists, and within high-level committees. This approach appears beneficial and applicable to other use cases, possibly with variations according to how critical the impacted business function is.

INITIAL FUNCTIONAL VALIDATION

The impact of an AI algorithm on the initial validation process should be defined as early as the design phase. Stakeholders involved should include technical staff from the design and validation teams, domain experts, and transverse committees concerned with the business processes in question.

5.3.3. Ongoing functional validation

Exploratory works on the AML-CFT topic, wherein an ML algorithm detects anomalies to be analysed by teams at levels 1 and 2, have shed light on how such an algorithm requires more sophisticated methods for ongoing review than traditional methods do. This includes continuously monitoring the proper calibration of the algorithm: volume of alerts raised to level 2, rate of false positives filtered out downstream, etc.

This upgrade of the ongoing validation process thus requires from the teams in charge:

- At a minimum, deploying and mastering tools for monitoring the operational behaviour (in real time if need be) of the algorithms.
- Building appropriate expertise and tradecraft so as to detect incidents upfront, and ideally to diagnose and remedy them as well.

ONGOING FUNCTIONAL VALIDATION

Ongoing functional validation of AI algorithms requires both dedicated tools (such as dashboards enabling the teams in charge to monitor their overall behavior) and closely interacting with the technical experts who designed them and performed the initial validation.

5.3.4. The case of internal risk model updates

Whenever AI is used in the construction of internal models, an essential consideration in the validation process is how to define triggering events for a model revalidation. In this regard, AI-based internal models differ from expert systems based on rules and various predefined configuration parameters, which have to be revalidated each time those parameters are proactively updated or are deemed obsolete: ML-based internal models become invalid following a major change in their input data. (It should however be noted that those models are not always devoid of predefined configuration parameters, such as a learning algorithm's hyper-parameters, in which case they must be subjected to the same treatment as traditional, rule-based models.)

In the particular case of internal risk models (referred to as "Basel models" in the banking sector), a model update policy defined by the banking institution clearly documents a number of criteria (such as a specific threshold being exceeded upon a parameter adjustment) which trigger the requirement to report the model update to the supervisory authority. This kind of parameter adjustment is typically decided and performed by the human experts in charge of the risk model, therefore one may ask what should become of those triggering events if the model is based on ML.

In fact, "classical" internal models assume that parameters are calibrated against the data, not unlike the training phase for an ML model. Besides, current regulation requires a governance framework which comprises the processes of back-testing and model parameter updating: parameter updates are often the consequence of back-testing results and thus caused by changes observed in the data.

The internal model update must be reported as soon as the induced change is deemed material by the institution, which should therefore define within its governance framework the process for evaluating the materiality of a change – whether the model uses ML or not.

Lastly, most actors in the banking sector opt for a scheduled calibration strategy; however some classical internal models update their parameters on a periodic, systematic basis (for example their volatility parameters in the case of a market model), also similarly to ML models. Therefore, from a model update policy standpoint, it appears that ML-based models can be treated like traditional internal models.

5.3.5. Technical validation

Technical expertise is required for AI validation, typically throughout the Data Science spectrum:

- Data Owner and Data Steward are respectively responsible for the governance and for the quality of data used by the algorithms.
- Data Engineers and Data Scientists are tasked with ensuring proper operational behaviour of software components which implement the algorithms.
- Lastly, in this context Data Analysts perform initial and ongoing validation of the algorithms' output.

The AI engineering stages to be covered by an adequate technical validation process include:

- Model selection, training and tuning.
- Continuous monitoring of the model (non-regression of the algorithm, absence of model drift, etc.).
- The higher-value task of detecting new data sources or variables to feed the algorithm.

TECHNICAL VALIDATION

Technical validation of AI algorithms introduced in business processes requires developing a broad in-house Data Science expertise, be it spread across departments (within an organizational structure built around multi-disciplinary profiles) or gathered within a specialized business unit.

This technical expertise should span the Data Science spectrum (from data engineering to state-of-the-art ML techniques) and be multi-tiered: generalist skills, financial sector specialization, and deep knowledge of business processes specific to the organization.

5.3.6. Management of AI-related risks

Internal control procedures are inevitably impacted by the use of AI and their evolution closely related to the generated risks, depending on the type of integration with business processes (automated vs. computer-aided decision-making) and on the nature of those risks (regulatory or legal, operational, financial).

Let us consider the example of AI used for assisting insurance claim processing: this downstream process within the value chain of insurance product distribution – which is not part of the exploratory works presented in this documented – has recently experienced an increased usage of ML. The typical use case in this context is to perform algorithmic filtering of incoming damage reports, so as to detect likely fraud cases or apply exclusion criteria. Cases having gone through those filters will result in a compensation offer (automatically generated in some cases), whereas rejected cases will be routed toward operators in charge of insurance claim files. The main challenge here is the accuracy of the claim management process:

- The insurance organization is exposed to financial risk if the rate of incoming claims resulting in a compensation offer unduly increases.
- Operational risk exists in case the volume of cases rejected by the algorithms increases to the point of overloading level-2 teams who need to review them.
- Lastly, compliance risk appears if the overload is such that the rate of disputed claims itself increases significantly.

Conversely, the main stake for the workshop on customer protection (section 8.5) is the explainability of the advice given, in order for the consumer who is offered an insurance product to be informed prior to making a decision.

Besides these various AI-related risks, a cross-cutting concern is the necessity for a dedicated safety mechanism (as part of a fall-back plan going all the way to business continuity planning) designed to remedy an incident, major malfunction or failure of an AI component:

- If the integration into initial business processes is sufficiently modular and robust, this safety mechanism may simply consist of falling back onto the initial process for the time period necessary to fix the failure.
- If the process has been more structurally impacted by the integration of AI, the safety mechanism will require more sophistication (and often proves more complex to implement as it also needs to be officially validated).

MANAGEMENT OF AI-RELATED RISKS

The nature of risks pertaining to AI's role in the process should be carefully identified: operational IT-related risk, financial risk, compliance risk, etc.

These risks should be included in the risk map required by governance principles, up to and including incident remediation and business continuity plans.

5.4. Security and outsourcing

Security of solutions relying on ML algorithms requires taking into account at least two types of risks rarely – it at all – encountered in traditional solutions: specific algorithmic risk (in terms of availability and integrity) and data processing risk. An additional consideration is the potential outsourcing of the design, implementation or exploitation of those solutions, which bears ML-specific security risks.

5.4.1. ML security

ML security challenges are similar to those of traditional IT systems: they typically pertain to confidentiality, integrity, and availability. Their treatment, although it should be tailored to their exact usage of ML, is in no way unique to the financial sector. This is even more so as the attack surface in finance is narrower than in other sectors: IT security in finance is usually a well-funded and mature area, furthermore exposure from things like open source code and use of public data has thus far tended to be more limited than elsewhere[14].

The way to make an ML model safe is different from the way in which a web service exposed through a REST API[15] – or the underlying data sources for that matter – can be secured. These three potential targets lie on three different architectural layers (while being mutually interwoven): respectively the model layer, the application layer, and the data layer.

A comprehensive description of the potential flaws of an ML model and of the means to remedy them is beyond the scope of this document. A categorization of the main possible attacks is however given in appendix 12.

[14] The reversion of this trend is however already visible, as many actors rely heavily (and sometimes exclusively) on open-source libraries and products implementing ML functionality so as to avoid "re-inventing the wheel", and as the use of so-called alternative data (collected from the web or from other publically available sources) becomes widespread among data-driven systems in finance.

[15] REST (Representational State Transfer) is an architectural method commonly used to build applications exposed on the web. A REST API (Application Programming Interface) is thus a simple, standard, easily secured method to design and deploy a web service.

ML SECURITY

Assessing the security of processes involving AI should take into account both classical IT security evaluation and the analysis of AI-specific risks and mitigation techniques.

In particular, any impact assessment, incident remediation plan or security audit mission should consider potential attacks against ML models, against their underlying data, and against predictive and decision-making systems as a whole.

5.4.2. Third-party and outsourcing risks

Financial actors rely on different types of third-party providers to develop their AI: design and implementation may be outsourced, off-the-shelf software products and services are now a common offering in AI, lastly hosting and administration of AI applications are often delegated to a cloud or hosting services provider.

Outsourcing-related risks

Risks classically generated by the outsourcing of software skills, design and implementation are particularly acute in AI.

Those risks are difficult to mitigate in practice if they have not been sufficiently anticipated. Hence a good practice prior to any outsourcing decision is to perform an *ex-ante* risk analysis taking into account the following risks.

Reversibility

As reported by the outsourcing guidelines published by European control authorities (EBA, 2019; EIOPA, 2020), reversibility of outsourced solutions constitutes a significant, non-AI-specific source of vulnerability within financial institutions today.

Using AI may even further exacerbate those concerns. Controlling the entire engineering workflow when it is outsourced requires mastering a variety of skills, including:

- Data Science skills pertaining to advanced ML techniques.
- Software design and architecture tradecraft related to complex systems with multiple integration points among components whose source code is not always open and well-documented.
- DevOps expertise in order to manage a heterogeneous infrastructure, which often combines dedicated hosting servers and cloud services.

Even when the entire skillset is available, as is often the case in large banking institutions, those skills may be scattered across departments whose technical teams are too “siloed” to be able to re-internalize what had been delivered (often as a monolith) by third-parties.

Outsourced AI development

As described in the workshop on probability of default (section 8.4), outsourcing the development of an AI component induces many changes in the business process. Resulting challenges are, among others:

- Developing and nurturing adequate in-house resources (human and technological) to validate code written outside the organization.
- Ensuring the delivered software is thoroughly documented as possible and periodically updated, so as to meet the criteria of internal control procedures.
- Planning for an audit of the deployed solution as early as in the AI design phase, which requires thinking ahead in terms of software architecture and integration capabilities.
- Deploying sophisticated (and themselves well-documented) explanatory methods in order to explain the results produced by a system whose development was outsourced.

Off-the-shelf AI software

The risks induced by an AI software acquisition strategy are similar to those resulting from outsourcing its development: dependency risk, non-reproducibility of results, lacking IT security by the software provider, product support deficiencies, and audit capability (assuming audit operations are relevant, i.e. when recurrently buying from the same provider).

The workshop on probability of default (section 8.4) has raised the issue of the dependency risk towards an AI solution provider: in that specific case, the risk is controlled insofar as the provider enables the customer to review all stages leading to the delivered ML model. It remains nonetheless the customer's responsibility to master the technologies involved in order to mitigate the dependency and vendor lock-in risks. Furthermore, caution is advised against acquiring software which does not sufficiently limit this dependency risk, for instance if the resulting model constitutes the only deliverable, to the exclusion of upstream stages in the engineering workflow allowing to rebuild the model – or alternatively, if that workflow is not adequately documented. This lack of information, both on how the ML model was designed and how it can be expected to behave, may lead both to operational risk and to difficulties in internal control and audit missions.

Lastly, it should be noted that a risk such as non-reproducibility of results is neither new nor AI-specific: traditional models which AI aims to replace often share this characteristic with outsourced AI solutions, especially if they rely on stochastic methods such as Monte Carlo simulations[16].

Cloud hosting

Service and application hosting on a public cloud has become an outsourcing scenario commonly encountered in the financial sector. To accompany this trend, an initial set of outsourcing guidelines have been published by European control authorities in banking (EBA, 2019) and in insurance (EIOPA, 2020).

Those guidelines cover more or less the same ground in both domains, namely: assessing how critical business processes are (and impact analysis), documentation requirements, duty to inform the supervisor, access and audit rights by the financial institution but also by the supervisor, IT security, the risks associated to data management, subcontracting, contingency planning (including business continuity plans), and the exit strategy out of the outsourcing agreement.

[16] Monte Carlo simulations are a class of optimization methods which relies on randomness (more precisely, repeated, computationally-intensive random sampling) to emulate the behaviour of an often deterministic process or model.

THIRD-PARTY AND OUTSOURCING RISKS

Any decision to outsource the design, implementation, hosting or operations of an AI system, or to use third-party products or services, must be preceded by an *ex-ante* risk analysis and take into account its results, especially with regard to reversibility.

Some governance principles pertaining to the third-party relation should also be observed:

- Ensuring proper documentation of deliverables and traceability of the process so as to enable auditing if necessary.
- Guaranteeing the financial institution access (technical, practical, but also legal with regard to IP – intellectual property – rights) to the source code and the ML models, even when development or hosting of the latter has been outsourced.
- Offering the same guarantee to the supervisor so as to enable audit missions which cover the AI systems, their source code, and the data they use.

5.5. Audit of AI algorithms

The AI evaluation principles previously exposed in the context of initial and ongoing validation remain valid for audit operations, be they performed by internal teams as part of periodic review or by the control authority as part of its supervisory missions. Thus, an auditor will need to consider the precise context in which the algorithm was developed and, in particular, the business processes into which it is integrated or which are impacted by it in one way or another.

Based on this context and on their objectives, auditors will need to consider the aforementioned trade-offs between the different evaluation criteria of AI algorithms, and the evaluation methods themselves will need to be suitable for self-learning systems. The audit team's AI skills must be sufficient to meet these requirements – as do those of the teams in charge of ongoing review.

5.5.1. A multi-pronged approach

A variety of situations can be encountered when evaluating an ML algorithm, due to the previously mentioned parameters (combination of algorithm type, end users and application scenario) and to the circumstances of the validation process itself (access to source code and underlying data may or may not be possible, technical resources may be available or not, etc.). The necessity to handle these different situations encourages the adoption of a multidimensional approach to the evaluation of ML algorithms: the one described in the following associates analytical and empirical methods.

Analytical evaluation

An AI algorithm can be characterized by its maximum explainability level, according to the four levels identified in section 3.4.4.

Levels 1 (observation) and 2 (justification) do not make the algorithm's internal behaviour intelligible: an explanation then cannot rely on the model architecture nor on the piecewise analysis of the algorithm at various levels of granularity. Levels 3 (approximation) and 4 (replication) on the other hand rely on a structural or detailed analysis of the algorithm – or more precisely analysis of its source code and of the resulting model.

If the organizational risk policy has adequately determined the explainability level required by the use cases of each AI algorithm, then audit missions pertaining to those algorithms should focus on high-stake algorithms which have logically been assigned a higher explainability level (3 or 4). In that case, the analytical evaluation is feasible, assuming a few prerequisites are met – the most important one being the accessibility of the source code, including its documentation.

In cases where an internal or external audit mission addresses an algorithm which has been assigned a lower explainability level (1 or 2), one of the first stages of the audit should consist of validating that this level is compatible with the types of risks and the compliance requirements of the business process in which the algorithm is integrated. The audit may then evaluate the algorithm and its impact on the efficiency of that process via the empirical methods described in the following section.

ANALYTICAL EVALUATION OF ML

When the stakes warrant it, appropriate analytical evaluation techniques and tools should be implemented as early as the algorithm design stage. Those methods may rely on information sheets describing the algorithm, the model, and the data used, and whenever possible on the analysis of the code and data themselves.

Empirical evaluation

The ML system is in that case treated as a “black box” and is evaluated from the outside, i.e. by observing its behaviour based on various input data. Several approaches are feasible, of which three are described in the following.

Post-modelling explanatory methods. These methods operate on pre-trained ML models and are categorized as global or local depending on whether they aim to explain a specific decision or the algorithm’s overall behaviour. A number of such methods are called “black-box” explanatory methods: they remain valid even when it is impossible to access the algorithm’s documentation or source code, and are therefore particularly suitable for algorithms whose maximum required explainability is level 1 or 2.

An auditor can also use post-modelling explanatory methods as a complement to any explanatory method implemented by the algorithm’s designer (such as those described in appendix 11.2). Besides, counterfactual explanations constitute a particularly interesting case of post-modelling explanatory method insofar as they can contribute to assessing that the appropriate data management principle described in this document has been followed, both in terms of regulatory compliance (specifically with respect to GDPR) and in terms of ethics and fairness. A non-comprehensive list of post-modelling explanatory methods is provided as appendix 11.3.

The workshops led by the ACPR with financial actors, also detailed as appendix, have shown those explanatory methods to be already in widespread use within the internal validation processes of ML algorithms, mostly as a way to assess their proper behaviour by means other than efficiency metrics. It seems logical that the same methods should also be made available to external actors tasked with evaluating those systems.

Benchmark datasets. This method consists of providing test data designed to stress-test the algorithm. That dataset may be either synthetically generated or composed of real-world data (anonymized if need be), or even hybrid (typically via a generative model which enables to augment an initial "bootstrap" data sample).

From a technical standpoint, any empirical evaluation of this kind requires dedicated Data Science resources, more specifically Data Engineering profiles who can build benchmark datasets and frameworks.

Challenger models. This method consists of providing a challenger model, whose predictions or decisions are to be compared against those produced by the model under evaluation.

A point of attention regarding this method is its practical feasibility: indeed, the development of challenger models requires allocating significant resources (human and material) and time to the task. Those constraints are also hardly compatible with audit missions as currently performed by the supervisor, which consist of analysing the properties of the model in place and of checking their consistency with respect to regulatory requirements. Thus they do not aim at building an alternative ML model[17]. The next section suggests a few ways to implement this type of empirical evaluation method – which is both an ambitious and a complex task.

Lastly, it should be noted that multiple empirical evaluation metrics are available and that their choice depends on the objectives, both when using benchmark datasets and for challenger models. Some metrics focus on efficacy (in order to assess algorithmic performance) whereas others analyse how particular segments of the population are treated (in order to detect discriminatory biases), others still investigate decisions made by the algorithm on a particular data point, etc.

[17] To put in perspective the effort required for building an alternative model, implementing a credit model for a banking institution typically involves tens of employees over a timespan of several years, even though its scope is limited to the organization's own data.

Empirical evaluation of ML

Empirical evaluation methods should be implemented as early as the AI algorithm design stage, and included in the quality assurance process of the resulting models (as part of the non-regression, functional, and integration tests).

Explanatory methods should be viewed as an essential tool for evaluating an ML model. They may be implemented at the design stage or operate on previously trained models, besides some methods apply to models that need to be evaluated as "black boxes". The choice of the most appropriate explanatory methods should take into account the algorithm type, the intended audience of the explanations, and the risk associated to the process.

Internal audit and supervisory missions may also use empirical evaluation methods such as benchmarking using their own test scenarios and datasets, or the comparison of challenger models to the models deployed within the organization. In order to facilitate those missions, it is recommended to design data models and algorithms to be as modular and well-documented as possible – which constitutes good software engineering practice anyway.

5.5.2. Challenges for the supervisor

The previously described multidimensional approach to evaluating AI algorithms requires the supervisory authority to adapt its tools, methods and data. Indeed, the analysis of an ML component differs from that of a procedural algorithm (and even more so of a process performed by human operators). It requires not only a certain level of expertise, but significant resources dedicated to building tools – such as challenger models – and maintain datasets which enable efficient control missions.

Besides, the variety of use cases for AI, and the variety of possible models for a given use case[18], require that the supervisor find a balance in the implementation of its evaluation methods: on the one hand adaptability is necessary to support the inevitable diversity of models encountered in different organizations, on the other hand only a sufficient level of formalism enables a systematic approach (e.g. to be able to plug a challenger model on the organization's data, and conversely to test any deployed model against a benchmark dataset).

Work on the tools

This line of work consists of building software and Data Science components in order to facilitate and accelerate supervisory missions.

Those tools should enable producing challenger models as previously described, in order to compare them against those provided by the supervised organizations. A specific obstacle for the supervisor is the dependency on heterogeneous data models across financial actors: the evaluation method based on challenger models implies that those models be able to ingest data according to a structure specific to each organization. The challenge is analogous at the output: for example, some of the transaction

[18] In the case of internal risk models, for instance, the diversity of models can be seen as a factor of prevention against herd behaviour and hence against systemic risk (see discussion paper *"Artificial intelligence: challenges for the financial sector"* published by the ACPR in December 2018).

monitoring models described in the workshop on AML-CFT (section 8.1) produce a categorical output (low/medium/high risk level) whereas others produce a numeric suspicion score.

Work on the data

This line of work consists of enabling the supervisory authority to access and process data from various open or closed sources (public data, regulatory data, supervisory mission reports, etc.) at different levels (national, European or international control authorities).

Those data should allow to build and maintain datasets for benchmarking the models deployed in the industry: the goals are to measure those models' performance, to assess their explainability on new kinds of data, to detect their temporal drifts, and so on.

The main challenge presented by the benchmark dataset approach is closely related to that described for challenger models: in the absence of standardization efforts, datasets must be produced according to a format and semantics aligned with those in place within the supervised organizations, which here also incurs an additional cost of technological and methodological adaptation.

Training

In order to enable and accompany this adaptation of the methods, tools and data available to the supervisor, appropriate training is a clear requirement, above all in the field of Data Science: such training may find its place at the supervision authority – as is considered within the ACPR – or in specialized institutions (for example pertaining to compliance in the banking sector).

6. Public consultation

Respondents are invited to illustrate their answers to the following questions using the use cases of AI – particularly of ML – implemented in their organization.

6.1. Context

QUESTION 1: EXPERIENCE WITH ML

- What kind of knowledge or experience do you possess regarding AI in general and ML in particular (R&D, Data Science, operational tradecraft, etc.)?
- If you are answering on behalf of a financial institution, what is the level of familiarity with AI within your personnel (both in technical and in business roles)?

QUESTION 2: IMPLEMENTATION OF ML *(QUESTION SOLELY FOR FINANCIAL CORPORATIONS)*

- What are the ML algorithms implemented in your organization?
- For each algorithm type, specify their use cases and the type of environment (development, pre-production, production)?
- For each use case, according to which criteria and evaluation methods has the algorithm been selected (raw performance, explainability/efficacy trade-off, etc.)?
- What are the respective roles of the teams involved in the design and implementation of ML algorithms in your organization (Data Scientist, software architects, project management, business experts, compliance officers, etc.)?

6.2. Explainability principle

On the basis of exploratory works conducted on three topics along with a broader investigation of AI in finance and of state-of-the-art ML in other domains, this document has outlined a number of expectations pertaining to the explainability of AI algorithms.

The relevance of those guidelines needs to be confirmed on several points, which are the object of the following questions.

QUESTION 3: DEFINITION OF THE EXPLANATION LEVELS

Are the four explanation levels emerging from this analysis (1: observation, 2: justification, 3: approximation, 4: replication) clearly defined? If not, indicate the points of misunderstanding.

Question 4: Adequacy of the explanation levels

Do those explanation levels appear to represent an adequate scale in the following senses:

- Do they span the entire spectrum of current and future applications of AI in finance, from full transparency all the way to algorithms operating as "black boxes"?
- Does the choice of four levels seem appropriate (if not, should there be fewer or more levels)?

Question 5: Practical examples of explanation levels

The table presented in section 3.4.1 suggests an appropriate explanation level for a few use cases of AI in the financial sector.

- How suitable are those suggested levels? If insufficiently, for what reason?
- Are those suggestions adapted to your own usage scenarios of AI (specify those scenarios)? If not, in what sense?

6.3. Performance principle

Question 6: Technical performance metrics

How do you view the technical performance metrics commonly used for ML (AUC or F1 score, GINI score, etc.), specifically:

- Their adequacy with respect to the various ML algorithms?
- The availability of methods to choose between those metrics?
- How those metrics are used (model validation, selection of its operating point, model drift detection, etc.)?

Question 7: Functional performance metrics

- Which functional metrics (KPI) seem relevant when evaluating an AI component? Do those metrics account for compliance requirements specific to the processes considered?
- Who should be responsible for defining functional metrics (technical or domain experts, with or without input from risk management and compliance teams)?

6.4. Stability principle

Question 8: Temporal drift of models

- What risks are, according to you, generated by the potential drift of models over time?
- What methods are or should be used to remedy those risks, or at least circumscribe them (out-of-time testing, alert triggering based on model drift detection, etc.)?

QUESTION 9: MODEL GENERALISATION

- What limits to the generalization power of ML models have been identified, whether in relation to overfitting or to intrinsic limits of the model?
- How can those limits be handled (out-of-sample testing, etc.)?

QUESTION 10: RETRAINING AS A SOURCE OF INSTABILITY

- Based on your experience, are model retraining phases (whether on a periodic or continuous basis) a source of model instability?
- What techniques are or could be used to limit this source of instability (non-regression testing with appropriate datasets, etc.)?

6.5. Appropriate data management principle

QUESTION 11: REGULATORY COMPLIANCE OF DATA MANAGEMENT

In your experience, which methods or techniques appear advisable in order to ensure compliance with various regulatory requirements relative to data management:

- The GDPR?
- Other cross-cutting regulations?
- Sector-specific regulations, such as the European IDD (Insurance Distribution Directive)?

Specify what stage(s) of the AI development process (design / training / prediction) involve these methods and techniques.

QUESTION 12: BIAS DETECTION AND MITIGATION

Which methods appear advisable in order to analyze biases in ML systems, for each of the following types of bias:

- Pre-existing biases in the input data fed to the ML models?
- Biases present in the algorithms themselves?
- Biases within the models produced by the algorithms – and in their decisions and predictions?

More precisely, which fairness metrics enable the identification of biases, for example those with a discriminatory nature?

Which methods can be used to mitigate the undesired biases thusly identified?

6.6. Integration in business processes

QUESTION 13: ROLE OF AI

- Which are or should be, according to you, the outlines of a method to assess the integration of AI components in business processes?
- What should such a method enable to evaluate: how critical the function of those components is, how disruptive they are with respect to the traditional process, what human-machine interactions are possible, etc.?
- What are your thoughts on maintaining "parallel" processes assigned to human operators so as to continuously evaluate and/or correct an algorithm's results?

QUESTION 14: AI ENGINEERING METHODOLOGY

- Should the engineering methodology used for AI differ from that used for traditional models, and more generally from standard software engineering practices? If so, in what way?
- How should, according to you, the ML model-building process take into account the integration of those models in business processes?

6.7. Internal control system

QUESTION 15: RISK MANAGEMENT

- How does the introduction of AI into business processes impact risk management: does it generate new risks or magnify pre-existing risks (specify the nature of those risks: operational or financial, legal, etc.)?
- Are new, AI-specific risk management methods called for (for example, calibration of ML models in order to limit the exposure to a given type of risk)?

QUESTION 16: FUNCTIONAL VALIDATION

- What should be the initial functional validation process of an ML model (i.e. prior to deployment in production)?
- Should functional validation be re-iterated when deploying a new version? Specify if the answer depends on the type of update (patch, improvement, etc.).
- How should ML components be continuously monitored for business risks?

QUESTION 17: INTERNAL MODEL UPDATE POLICY *(INTERNAL RISK MODELS)*

- On what conditions may, according to you, ML algorithms be used within "Basel models" in the banking sector, and within internal models in the insurance sector?
- How should an organization's internal model update policy take into account the use of ML in its internal models?

QUESTION 18: TECHNICAL VALIDATION

- What should be the initial technical validation process of an ML model (i.e. prior to deployment in production)?
- What technical indicators and methods should be used to continuously monitor ML components deployed in production?

6.8. Security and outsourcing

QUESTION 19: OUTSOURCING

Does the use of AI generate specific challenges or risks when its development, hosting or administration are outsourced? If so, which ones?

QUESTION 20 : SECURITY

- What is the impact of using ML on IT security?
- Which types of attack against ML models (causative attacks, surrogate model attacks, adversarial attacks, etc.) appear the most important to you, both in terms of occurrence likelihood and in terms of damage inflicted in case of success? Specify according to the type of ML model, the use case, and the environment (dedicated hosting servers or cloud services, etc.).

6.9. Multi-pronged approach to evaluation

This document suggests implementing a multidimensional approach for auditing processes using AI. The following questions aim to further define this approach.

QUESTION 21: ANALYTICAL EVALUATION

- Which of the following elements are available for evaluating an AI algorithm in the relevant organizations: the source code? Its documentation? The resulting models? The training and validation data? Specify if the answer depends on the algorithm type, the use case involved, or the context of the evaluation (internal validation, external audit, etc.).
- Do you use standardized documentation frameworks such as information sheets describing the algorithm, the model, or the data used?

QUESTION 22: EMPIRICAL EVALUATION

- Which of the empirical evaluation methods suggested in section 5.5.1 (benchmark datasets or challenger models) seems more appropriate in your opinion?
- Is the architecture of data processing workflows and AI systems within the relevant organizations sufficiently modular and robust to enable this kind of functional testing at the data or model level?
- Are the data format and schema sufficiently standardized (or flexible) to support a data benchmarking method without incurring data integration costs by the supervisor?
- Analogously, are they sufficiently documented and transparent to support the integration of challenger models developed by the supervisor, without this approach being rendered unrealistic by an information asymmetry?

QUESTION 23 : EXPLANATORY METHODS

- Which explanatory methods (cf. appendix 11) are currently implemented among the use cases of AI to your knowledge?
- Do you know of any explanatory method for AI other than those described in this document? If so, which ones? Have they already been implemented and deployed?
- Does the most appropriate explanatory method to use depend on the algorithm type?
- Does it depend on the intended recipients of the explanation, and if so, in what way?
- Does it depend on the level of risk associated with the business process, and if so, in what way?

Appendices

7. Technical scope

AI is an extremely broad field whose definition – based on academic work and industry practices – evolves quickly over time. Within this discussion document, AI is considered solely in its embodiments relevant to the financial sector, both in their current form and in their likely evolutions over the near- to medium-term horizon.

7.1. ML vs. AI

The scope of this document is restricted to ML (Machine Learning), which happens to be probably the most intensely studied field within AI. Other forms of AI are not taken into consideration: robotics, game theory, optimization under constraints, multi-agent systems, knowledge representation and reasoning, or planning automation.

Among the ML methods used in the financial sector and considered in this document, the following categories should be mentioned (without any comprehensiveness):

- Unsupervised learning methods (in particular clustering techniques), which are commonly used in fraud and anomaly detection scenarios.
- Predictive models which may be called "traditional", such as decision trees and logistic or linear regressions.
- More sophisticated yet also commonly implemented models such as decision-tree based ensemble methods (Random Forests, Gradient Tree Boosting, etc.).
- NLP (Natural Language Processing), used to classify and analyse all kinds of text data.
- Deep Learning (deep neural networks), used in various use cases including CV (Computer Vision) where they particularly shine – although a less prominent use case in finance than in other sectors.

7.2. Models vs. algorithms

Another key point of terminology is the distinction between an AI algorithm and the model produced by that algorithm. An ML algorithm (AI being, as indicated above, the field of AI considered in this document) is an executable procedure represented as software code, just like any algorithm. Its specificity with respect to other types of algorithms is to operate on input data (above all training data but also validation data) and to produce an ML model as output. That model is, generally speaking, itself composed of a predictive algorithm and of model data. The predictive algorithm is typically an optimization procedure which minimizes an error metric for the model on training data.

A few examples shall illustrate the relations between ML models and algorithms:

- A linear regression algorithm produces a model composed of a vector of weights.
- A decision-tree construction algorithm produces a model which is a tree whose internal nodes are logical conditions involving predictor variables, and whose leaves are predicted values.
- A neural network algorithm (based e.g. on a back-propagation method and a gradient descent algorithm) produces a model which is a graph structure whose nodes are weight vectors.

The terms model and algorithm are sometimes used interchangeably within the present document when the context is unambiguous, or when the meaning refers both to the model building process realized by the algorithm and to the prediction process realized by the already-built model.

8. Detailed description of the exploratory works

This appendix presents, for each exploratory work on the three topics selected:

- A description of the purpose and relevance of the exercise.
- The objectives of the algorithm presented by the financial actor involved.
- A few technical details on the method and the implementation.
- The validation process adopted by the actor.
- The governance issues raised by the introduction of AI into the business process.
- The evaluation methods used and their implications, according to the four evaluation principles exposed in this document (appropriate data management, performance, stability, explainability).
- The engineering methodology used to develop the AI system in question.

The following sections do not in any way constitute an evaluation of the algorithms studied during the exploratory works, nor of the business processes in which they are used[19]. Their goal is to provide contextual, factual information to the reader, so as to shed light on the lessons drawn by the ACPR in this discussion document.

8.1. Topic 1: Anti-money laundering and combating the financing of terrorism (AML-CFT)

8.1.1. Regulatory context

Current AML-CFT regulation requires financial institutions to implement risk management procedures enabling them to detect PEPs (Politically exposed persons), the transactions involving individuals tied to a high-risk country listed by the FATF (Financial Action Task Force) or the European Commission, as well as the transactions, which are incoherent or anomalous with respect to the organization's knowledge of its customers, and may result in a SAR (suspicious activity report, or equivalently suspicious transaction report).

European and national regulations pertaining to freezing of assets also require financial institutions to set up a unit dedicated to implementing the relevant measures – which include, in addition to asset freezing, the prohibition of making funds available.

Those regulations do not require using a computer system to do so, but in practice most organizations use software processes due to their size and their activity volume.

Lastly, those regulations do not contain any provision specific to the use of AI.

8.1.2. Purpose of the exercise

The objectives of the primary AML-CFT workshop, augmented with a secondary workshop, were the following:

- Understanding the potential use cases of AI in AML-CFT.
- Gaining familiarity with the underlying AI techniques.

[19] The call for applications published in March 2019 stated that the works envisioned were in no way related to the ACPR's supervisory procedures.

- Thinking about possible adjustments of supervisory processes in view of controlling AI-driven AML-CFT processes.

8.1.3. Objectives of the algorithm

The main project studied within this topic consists of introducing ML models to aid the filtering of transactional messages – in other words, design algorithms which can assist agents tasked with distinguishing, among the list of alerts raised by a rule-based third-party monitoring tool, the false positives from the transactions concerning individuals who actually are on embargo or sanction lists.

In the process prior to introducing AI, operators review alerts issued by the screening mechanism in order to determine whether they are physical or moral persons targeted by restrictive measures. These operators are organized according to two levels. A level-1 team is in charge of the initial alert processing, on the basis of a decision-making matrix. The alerts which are not resolved at level 1 are escalated to level-2 teams, which are authorized to release the payment, reject it, or file a homonymy case with the administrative authority responsible for the freezing of assets.

The role of the ML model developed is to assist this decision-making process and to route the transactional messages to the appropriate level based on their relevance, i.e. the more sensitive messages will be directly processed by level 2, which aims at streamlining and securing the overall process. Level-1 teams, no longer in charge of the initial processing of some of the alerts, will then be able to absorb a volume increase. This model, developed by the participant to this workshop, was dubbed TPA (True Positive Acceleration).

8.1.4. Technical details

The ML algorithm is based on a neural network which is fed features with varying levels of complexity: message characteristics, phonetic distances between strings, address components (using NER, i.e. named entity recognition), and semantic analysis of free-form text. Those variables are extracted from transactional messages by the filtering tool and do not contain any personal data (unlike the original messages).

AI contributes to rationalizing the filtering process. Indeed, by quickly and efficiently discriminating between heaps of voluminous messages not only frees up the analysts who can focus on tasks with higher added-value. The analysis of results produced by the AI also gives them higher accuracy in their daily job, since the risk forecasting process gains in precision as the volume of data analysed grows over time. A reduced amount of routine, repetitive tasks, along with the opportunity to partake in engaging strategic works, should also contribute to employee retention.

Lastly, the situation can also be considered where AI directly contributes to improving the decision-making of human analysts by performing a *post-hoc* analysis of the abandoned or escalated alerts, so as to give them a means to adjust their decisions on future alerts.

8.1.5. Validation process

The starting point for the participant to this workshop was to build on existing validation methods used for risk management models, which could be relevant to internal control procedures.

The usual frameworks for such risk management models are organized around a model validation team and a model update team. Those two teams are mutually independent: an independent review tends to increase the algorithm's efficacy and to reduce its operational risk.

The goal is to perform a formal validation once a year and each time the model undergoes significant change. Meanwhile, machine or human expert systems might use rule-based procedures in order to build a reference dataset, which can then operate as a benchmark against which to compare the model under development, so as to identify cases where AI-driven decisions deviate from expected norms.

The peculiarity of validation processes for ML is the lifecycle of the models:

- On the one hand, the integration of the ML component in the business process should be performed once, according to validation methods in line with the organization's governance framework.
- On the other hand, the statistical validation of the model should be consistent with the first kind of validation, and be repeated over time – ideally on a continuous basis.

In other words, the notion of *a priori* validation should be re-examined, since shorter validation cycles are necessary, which makes the dichotomy between initial validation and ongoing review less relevant in the case of AI algorithms.

At any rate, the validation process should be proportional to the risks, in particular in terms of regulatory compliance.

8.1.6. Governance issues

The governance schema chosen by the actor was to ensure a two-fold human role in the monitoring of the algorithm: level-2 analysts are tasked to authorize or reject transactions, but also to guarantee the algorithm's proper behaviour, while level-1 analysts also annotate transactions in parallel with the algorithm, which increases the amount of additional training data available to it. The latter approach has not been retained by all actors who introduced AI into their AML-CFT filtering workflows (cf. section 8.1.9), however it enables to validate the performance and stability of the algorithm over time, even in the presence of major changes in transaction profiles.

In terms of operational risk, a point of attention is the significant decrease of the level-1 workload (on the order of 10%) due to the introduction of the ML model. It is necessary to anticipate the operational risk that would result either from an interruption of the algorithm's operation or from a more general system failure: this risk is critical because of its potential ramifications, given that the AI component contributes here to a performance obligation. In particular, the organization needs to ensure that the level-1 validation teams remain capable of absorbing, if necessary, the entirety of incoming transactions without degrading the quality of service provided.

8.1.7. Evaluation methods and their implications

Explainability

The explainability requirements of the algorithm are different from the other workshops (which pertain to credit granting models and to the construction of an insurance product).

Indeed, there is no requirement to motivate the decisions made by the algorithm which impact an individual. Checking the relevance of an alert raised by the algorithm is also relatively simple for an analyst: in order to do an efficient job of comparing the alert to a sanction list, the operator does not need to know the reasons why the alert was triggered.

The most important benefit of explainability in this case is its business value: it facilitates the analysis of the patterns of filtering behaviour captured by the algorithm (which also constitute its training data). This assistance in understanding the operations performed by a human analyst is an additional help for the ongoing review of the algorithm's efficacy, which is a key asset in a domain subjected to a performance obligation (otherwise put, false negatives are very costly and should be reduced to a strict minimum).

Performance

The statistical performance of the predictive model, along with its operational impact on the alert processing workflow, have been evaluated with the following observations:

- Statistically speaking, the model exhibits a slight overfitting, which however does not appear to induce any functional risk given how the algorithm is integrated in the overall process: in the worst case, that process will not be automatically accelerated, nevertheless they will be adequately processed by level-2 teams if need be.
- The algorithm's impact on the business process manifests itself by a significant decrease of the volume of alerts to be processed by level-1 teams, and by a marginal increase of the volume of alerts to be processed by level-2 teams due to the improved recall of the model.

Stability

The model's behaviour appears to be stable over time, insofar as the relative impact of the acceleration of message processing on the workload managed by level-1 and level-2 teams is itself stable over time.

However in this usage scenario, data quality and comprehensiveness are essential, and their "freshness" is necessary to ensuring that the model which relies on them is operating properly. Two approaches can be used to this aim:

- Making temporality explicit in the algorithm, since it plays a key role in the semantics of data: in particular, datasets used in AML-CFT should be periodically reviewed in order to take into account new methods used by malicious individuals.
- Building generic, time-independent variables: for instance, instead of using a "country" variable, use "country belongs to a given sanction list" which is a time-invariant feature related to the issue considered.

Appropriate data management

This project directly stems from the compliance department, however as previously indicated, a specificity introduced by the use of ML is that the responsibility for the validation process, in addition to the compliance team, also lies on domain experts and on technical experts.

8.1.8. AI engineering methodology

The project was undertaken according to an agile methodology, and was at the time of the workshop still at the experimental stage. As suggested in section 8.1.5, it appears sensible at this stage not to demand – even on as sensitive a topic as AML-CFT – an excessively broad or cumbersome validation process, which would involve other departments and hinder its deployment in production.

8.1.9. Secondary workshop

A secondary workshop on the AML-CFT topic was conducted with another banking group[20]. This section only summarizes the noteworthy differences with the primary workshop.

Objectives of the algorithm

The business process in which ML was integrated in this case is the filtering of transactional messages, not to screen them against sanction lists (which in the primary workshop led to a potential rejection of the payment or the assets being frozen), but to detect suspicious transactions and, when appropriate, yield a suspicious activity report (SAR). This function is performed by enterprise software specialized in filtering financial transactions, which uses preconfigured business rules: those rules are executed on each transaction to produce a suspicion score, which is then used to route transactions above a threshold toward teams tasked with the analysis of alerts. Following the standard, those teams are broken down into levels 1 and 2: alerts above a first threshold are directed to level 1 (at the level of the branch offices) while those above a second, higher threshold are directed to level 2 (the Tracfin[21] correspondents of the banking group).

In the new approach, an ML model is trained on a training dataset composed of 50% manually issued alerts which have been validated and 50% alerts generated by the rule-based software. It should be noted that for a significant portion of the manually issued alerts, the suspicion score produced by the business rules is zero.

The integration of ML into the process differs from the primary AML-CFT workshop in that the ML model is here introduced as a complement to the enterprise software, with the following features:

- The function of the ML model in the primary workshop was to escalate some of the alerts trigged by business rules from level 1 to level 2. In this case, the ML model produces additional alerts which are sent directly to level 2. The AI thus follows a parallel workflow, and not a serial one where the execution of business rules would be followed by ML prediction. Thus, rather than a classifier for previously-raised alerts, the banking group has deployed a detection tool for validated alerts which is applicable to the entire transaction flow.
- Also, a filter has been introduced so that, when a transaction is assigned a high suspicion score by the ML model, an alert will only be generated if no alert was raised by the business rule engine on the same customer within the three preceding months. In other words, an alert triggered by the ML corresponds to a customer which has been given a high score while having stayed below the rule engine's detection threshold for a while.
- Lastly, contrary to the business rule engine, the ML model takes into account information beyond transactional data: statistical features of the transactions are combined with static variables (either direct measures such as duration of the customer relationship or asset value, or constructed variables such as the types of products and contracts) on a sliding time window.

[20] This workshop is presented as a secondary study as it was conducted belatedly, furthermore the use case for AI and its technical implementation are relatively similar to the first workshop.

[21] Tracfin ("*Traitement du renseignement et action contre les circuits financiers clandestins*") is a service of the French Ministry of Finances in charge of enforce AML regulation and coordinating its application.

Governance issues

Contrary to the primary AML-CFT workshop, level-1 teams do not annotate the transactions in parallel with the ML model so as to detect false negatives: this is because according to the organization's AI team, any relevant sample (i.e. having a sufficient number of false negatives) would be too large. Two methods could be considered for analysing false negatives – namely either lowering the alert triggering threshold or systematically sending the n most suspicious cases for review –, both of which would likely induce an excessive additional workload for the operators. Besides, some false positives are simply due to non-observable variables.

It should be noted that this way of introducing ML into the business process (i.e. as a complement to the enterprise software), along with the routing of ML-generated alerts to level 2, result in additional workload for level-2 teams. This is why a new AI project has been initiated with the goal of routing certain alerts raised by the enterprise software from level 2 to level 1, in order to reduce that extra workload.

Also in relation with those changes in the business process, the banking group has decided to structure the organization of its AML-CFT expertise around "dual skillsets", i.e. employees who master both ML (including data management issues) and possess business experience (including in risk management).

Those different governance choices between both AML-CFT workshops are particularly interesting: each option is probably suitable for its particular context, and the feedbacks gathered around both projects will likely provide valuable know-how regarding the possible trade-offs between the predictive power of an ML model, its temporal stability, and the workload dedicated to manual annotation of data.

Explainability

Explainability requirements are aimed at different types of users. A joint effort within the banking group, involving technical teams, the compliance department and IT people, led to proposing explainability forms adapted to what each user type wishes to observe and in what context (in line with the approach described in section "Recipients of the explanation"):

- Technical teams (in particular Data Scientists) rely on the explanations during the model construction phase – not for continuous monitoring. SHAP (*Shapley Additive Explanations*) values are the explanation form used in that case to understand the decision made on a particular transaction.
- Compliance experts use explanations to support their decision to abandon or validate an alert. Workshops were organized with these users in order to better define their needs (as simple tabular representations of SHAP values were quickly deemed inadequate). This led to the development of a GUI (graphical user interface) showing explanations which are still based on SHAP values but easier to interpret and more actionable.
- Lastly, the banking group also aims to provide relevant explanations to internal or external auditors, including (as a complement to both previous explanation forms) a documentation ensuring proper intelligibility of the algorithm.

Performance

The main performance indicator is the rate of alerts generated by the detection system which result in a SAR. The introduction of ML according to the aforementioned architecture enabled the doubling of this indicator.

Stability

A monitoring tool has been implemented as early as the initial deployment of the ML model, so as to detect any operational anomaly or model drift. That tool periodically checks several indicators characterizing the model, the input data, the output score distribution, etc.

Technical teams indicated during the workshop that it was still too early to determine whether drifts of the ML model were more or less frequent than the need to reconfigure the enterprise software. Updating the ML model would nevertheless be simpler than updating the parameters of the business rule engine for several reasons: it is a simple retraining phase without addition of new features, it is also fully automatable, and the entirety of model parameters are adjusted without any manual intervention. Besides, the ML model update – from retraining to deployment to production – would not take longer than 2 to 3 days, which is significantly less than a reconfiguration of the enterprise software.

8.2. Topic 2: Internal models in banking and insurance

The second topic for the exploratory works conducted by the ACPR pertained to internal risk and capital requirements models. In fact, candidates on this topic suggested to study use cases in a slightly different domain.

As a consequence, this topic pivoted toward risk credit modelling, considering both granted to individuals and to businesses. It consisted of two distinct workshops:

- A workshop focusing on credit granting models: those models usually compute a credit score. The participant to this workshop is a banking group.
- Another workshop relative to so-called behavioural credit models: those models aim to estimate a probability of default on a given time horizon for a current credit. The participant to this workshop is a large consulting firm which provides to banking organizations an ML model construction platform.

8.2.1. Regulatory context

Both workshops shared the following initial observations:

- Classical internal models are generally relatively easy to audit but perform poorly. More advanced or more complex models should provide a performance improvement, albeit at the cost of explainability.
- Regulatory requirements are identified as hindrances to the implementation of innovative algorithms, especially those based on ML: such requirements pertain to stability of the resulting models, to their auditability, but also to the transparency and explainability of the algorithms.
- Additional challenges related to personal data protection, along with limitations inherent to the data (in terms of access or completeness, for example), make it challenging to analyse correlations among multiple variables characterizing customers and their behaviour.

8.3. Workshop on credit scoring

8.3.1. Purpose of the exercise

The banking group in question has implemented methodological guidelines for credit scoring models.

The workshop aimed to explain how credit modelling teams took into account those guidelines – which had been defined and refined over the course of many years – so as to build models in accordance.

8.3.2. Objectives of the algorithm

This workshop involved the analysis of several credit scoring models, all of which answered a dual objective:

- To reduce the dependency toward third-party data providers (such as Credit Bureau) by integrating additional internal data sources into the algorithms: for instance, behavioural data in addition to Credit Bureau scores and to traditional internal data such as credit history.
- A more classical objective was to improve the discriminating power of credit scoring models.

The three models studies were respectively about credit for enterprises, credit for the purchase of used vehicles, and credit for household equipment purchases.

The Household Equipment model is described in further detail in this section. The other two models present similar issues, both at a functional and at a technical level. The business objective of the Household Equipment model is to make a decision on the credit request within 5 minutes.

8.3.3. Technical details

The project relies on the following data sources:

- Data on credit applications
    - Individual data on the applicant (and co- applicant when appropriate)
    - Information on the product (amount, credit terms, etc.)
- Data on contractual risk
    - Data used for computing default states
    - Data used for computing behavioural variables
- External data
    - Credit Bureau scores
    - Data from central banks

Data Scientists met during the workshop insist specifically on the importance of enriching internal data (which is typically the only kind used in such projects) using external data: the latter will be of various types (text, time series, etc.) and sometimes collected from open data sources (obtained via web scraping). The strength of ML lies not only in using novel algorithms, but also in leveraging such data sources - often called “alternative” data sources.

Most models implemented by the teams decided to use a Gradient Tree Boosting algorithm (or variants thereof) after comparing it to other algorithms commonly used in the organization (in particular, SVMs were too demanding in computing resources, and neural networks were deemed unsuitable for this use case).

8.3.4. Validation process

The validation process within the banking group for any credit granting model developed using ML prior to its deployment in production (whether a new model or a patch on an already-deployed model) is as follows:

- Credit teams who designed the model (usually located in the same country, or centralized teams in cases where sufficient Data Science resources are not available at local entities) send the Validation team a dossier comprising a technical documentation along with the entire source code.
- The Validation team inspects the documentation (conceptual validation) and re-runs the model generation code (training, test, validation) in order to verify its results and to bring a critical look on the methods used. This is only possible because the Validation team possesses all necessary skills to evaluate the model according to the principles described in this document (data management, performance, stability, explainability).
- For certain entities of the banking group, credit granting models are used in Basel models (i.e. internal risk models in the banking sector): in such cases, the Validation team presents the model to the group's Risk Committee in order to get the strategy choice approved (e.g. constant risk, decreased risk, hybrid strategy).
- When appropriate, the dossier – once validated at the group level – is sent to the ECB for validation of prudential models.

The validation process thus comprises conceptual phases but also applied phases.

8.3.5. Governance issues

This workshop described a scenario where an ML component is introduced as a computer aid to a decision-making process (and not as a fully automated process). Indeed, the component is part of a multi-step process:

1. Execution of business rules (related to age, filters, over-indebtedness) previously defined by domain experts jointly with the Validation team.
2. Automatic computation of the credit score (which is given a lower weight than business rules in the overall decision-making process).
3. Possible intervention by a human agent, who can override the decision, both in cases of a high score (credit granted by the system) and in cases of a score below the threshold.

8.3.6. Evaluation methods and their implications

Explainability

There are multiple objectives for explanations in this use case:

- Model designers need to guarantee the proper behaviour of the algorithm and to facilitate the validation process.
- Explanations are also aimed at the teams responsible for continuously monitoring the system.
- Lastly, they will in the future be useful to agents who need to understand a negative result produced by the algorithm before making a decision, i.e. either confirming the credit denial or granting the credit through a manual override.

The SHAP method was retained for the three situations (LIME was also evaluated), for the following reasons:

- It enables both global explainability (i.e. which type of information weighs on the model's decisions) and local explainability (i.e. which values taken by a specific data point impact the decision positively or negatively).
- The form of explanation provided by SHAP has been deemed by users to be the most analogous to the traditional (logistic regression) model.
- The method was easy to implement in each of the three situations.

A counterfactual explanatory method (cf. section 11.3.1) is however also being considered: it would likely require a significant amount of UI work, especially if a large amount of information needs to be presented to users. Besides, the explanation should be as intuitive as possible, which is not straightforward in cases where the underlying decision tree has been split on criteria which are not quite logical (e.g. "`age < 23.5 years`").

Performance

The main methods and metrics retained to evaluate the model performance are the confusion matrix or F1 score to assess recall and precision, the GINI score to evaluate its discriminating power, and the Kappa coefficient for comparing the old and new scoring models.

In particular, a GINI threshold is defined by the guidelines implemented throughout the organization, both for all credit models (the current status being that this threshold is achievable for most re-designed models except on certain population segments such as younger age groups) and for all regulatory models (with a higher threshold in that case).

The GINI gain obtained when going from traditional scoring models to the ML model produced by Gradient Tree Boosting is rather small (a few percentage points) in the case examined during the workshop, i.e. the household equipment model. Nevertheless it can reach up to 23 percentage points in some models developed by the team, namely those which initially had a low discriminating power. Furthermore, even a seemingly marginal GINI gain generally represents a significant decrease in the key business metric in this case, namely the expected credit loss.

Stability

The main stability metric retained in this project is based on cross-validation results (namely some checks on the standard deviation over the different folds).

Several indicators are also monitored:

- Mutation rate of the population (using the Population Stability Index).
- Evolution of the portfolio profile (credit application rate, acceptance rate, number of defaults over the previous three months), in accordance with the monitoring practices described in section 3.3.
- Evolution of business performance metrics.

In case an alert is raised on those indicators, an analysis if performed in order to find probable causes for the corresponding statistical anomalies, and a remediation plan is produced, which may in some cases include a model redesign.

Due to lack of hindsight on the operation of the new model thus far (which is done in parallel with the traditional model still used in production), the teams were not able to estimate its stability nor its appropriate update frequency.

8.3.7. AI engineering methodology

The credit granting models developed by the teams are not yet in production. A method to analyse corporate credit risk, however, has been implemented and deployed: it leverages (mostly open) data in order to estimate a company's default risk.

8.4. Workshop on probability of default

8.4.1. Purpose of the exercise

A workshop was conducted with the Credit department of a consulting firm, who offers its clients from the financial sector an ML engineering solution which is applicable to building models to estimate probabilities of default. This workshop was quite complementary to the previous one which focused on credit scoring models insofar as it relates to a generic, fully externalized AI solution. It thus represents an interesting example of the adoption by a financial actor of an ML product developed by a third-party.

The solution offered by the consulting firm is not an off-the-shelf product operating as a black box, but a toolbox which enables to design and build a model while maintaining a constant interaction between the solution provider and the customer. In practice, the resulting model is a hybrid one, partly based on advanced ML algorithms during the design phase but then translated into simple and explainable algorithms for the deployment phase. This choice appears to have been motivated by the necessity to deliver a well-documented model, along with an audit track.

The solution as currently available is designed to support credit scoring and probability of default models, however the solution provider is working on applying a similar approach to internal risk models, namely leveraging ML to yield corrections and improvements to currently used models in the form of business rules.

8.4.2. Objectives of the algorithm

The main objectives of the project were the following:

- Increasing the performance of the models used for decision-making. In particular, an improved risk discrimination through the identification of non-linear effects between risk factors, an improved classification of individuals, and a faster identification of changes in the underlying risk portfolio.
- Improving data quality through the use of quality assessment and improvement techniques.
- Refining the estimation of regulatory capital requirements through the use of more accurate models.
- Increasing the transparency and auditability of the models.

Data availability is an essential issue in this case, since the volume of data that can be exploited varies greatly with the use case: few data points for consumer credit, far more for housing credit.

8.4.3. Technical details

The main stages of the nominal behaviour of the solution are commonly encountered when adopting advanced ML models, with the exception of the last one which makes the approach original. These stages are:

1. Data quality control and data preparation prior to modelling
2. Construction of a reference model (of a “traditional” type), in practice a logistic regression.
3. Construction of a challenger model (of an “advanced” type): more sophisticated, supervised ML algorithms are used, typically random forests or neural networks.

4. Identification of the margin of improvement of the reference model: in the use case considered, 80% of the prediction error can be attributed to 20% of the population, thus the goal is to identify population segments which are incorrectly classified by the reference model.
5. Visual explanation of the decisions made by the ML model: the methods used are classical ones (SHAP, LIME).
6. Extraction of simple, auditable business rules which explain the performance gap between the ML model and the reference model: to this aim, population segments which are incorrectly classified by the reference model are automatically identified, then business rules are extracted by a domain expert (typically from risk management) so as to reduce as much as possible that performance gap.
7. Definition of the final hybrid model, as a combination of the reference model and business rules.

The solution is offered as "managed services": besides the hybrid-model-building workflow described above, an information sharing platform enables the customer to review the entire design process independently from the execution of that workflow.

To some extent, the model building approach adopted here relies on challenger models mentioned in this document as a possible audit method (cf. section 5.5): several hundreds of model exemplars are compared against one another, then the best one is retained, following which the system will minimize the performance gap between the reference model and that "top challenger". In a nutshell, the strategy is to try to replicate the performance of the best challenger models while remaining inside a more controlled operational framework – which is guaranteed by combining an intrinsically explainable model (logistic regression) with a limited number of business rules.

The creators of the ML platform examined insist that the choice of avoiding "pure" ML models was made early on in the project, firstly because ML is notoriously difficult to implement in this type of scenario, secondly because such a model would hide behaviour inherent to the population considered, such as the transition of individuals across population segments over time – which is essentially observed in any credit model.

8.4.4. Validation process

Initial functional validation relies on the documentation of the algorithm and on a presentation of its results. It is performed by the solution provider in support of the customer, in an iterative mode which is more specifically focused on aforementioned stages 4 to 7 (i.e. from the identification of margins of improvement to the definition of the resulting hybrid model).

As for continuous functional validation, it is similar to back-testing which is usually performed for credit models, except that frequent monitoring of population segments impacted by business rules is required, the goal being to anticipate the detection of model biases. Interestingly, back-testing results are presented to the Risk Committee in order to assess the relevance of a model adjustment.

8.4.5. Governance issues

The solution design, which ultimately consists of tuning a reference model via business rules (stages 6 and 7), aims to make it compatible with governance frameworks common to most traditional models. In particular, the hybrid model designed by the consulting firm can be assimilated to the classical behaviour of credit granting models, which follows an analogous business process: a regression model

– comparable to IRB (Internal Ratings-Based Approach) models – is first executed, then an override (similar to the "notching" practiced by credit rating agencies) can be applied by a human agent if a weakness has been identified in the model's output. Besides its compatibility with a tried-and-tested governance framework, the benefit expected from the approach is a high model explainability (cf. next section). The choice of a hybrid model has also been made for various operational reasons: easier implementation, stability and robustness.

Another governance issue raised by this use case is however relatively common, namely the outsourcing of the model design and implementation, and also of its maintenance.

8.4.6. Evaluation methods and their implications

Explainability

By choosing a hybrid model based on decision rules, the solution provider has put the emphasis[22] on the generation of convincing explanations – local and global – intended both for users and for governance bodies.

For instance, an essential explainability criterion is that the aforementioned overrides of the model decisions must be motivated. In particular, a user of the model (typically an account manager) must understand why the model produced a given score. Besides, as explained in section 5.2.3, the intervention of a human agent introduces a risk of "explanatory bias" with respect to the more objective result provided by the model. In the hybrid model, business rules have been pre-selected by the algorithm: in essence, the model is first optimized as a logistic regression, then the addition of business rules aims to optimize the resulting hybrid model – in both cases in terms of global performance.

As for local explainability, the use of SHAP enables to provide the reasons for a particular score given by the logistic regression model. An explanation of the decision made by the overall hybrid model in turn consists of augmenting those SHAP values by the motives of any overrides made by the business rules. Those motives are quite simply the membership of the individual considered in one or more population segments on which the model's predictive performance had been optimized by the algorithm.

Performance

The following metrics, which combine predictive performance and business efficacy (cf. section 3.2), are used to assess the relevance of the overall model construction workflow:

- As predictive performance metric, the GINI score gain is used (typically on the order of 5% in the cases studied).
- Two business performance indicators are computed: the gain in terms of returns measured while keeping the risk appetite constant (around 50%) and the reduction in expected loss (which is a standard computation in internal risk models).

[22] This concern is also evident in certain technical choices: for example a genetic algorithm was picked for hyper-parameter tuning rather than e.g. a Bayesian optimization method, because it was deemed easier to explain even to laypeople while offering comparable performance.

Furthermore, the replicability of the model has been studied: initial runs experienced a problematic lack of reproducibility, which was later solved.

Stability

Firstly, this workshop illustrated the observation made in section 3.3.1, namely that the temporal drift of a predictive model may in most cases be due to a significant change in input data, without even considering the impact of the ML algorithm. Thus in the case of credit models, structural modifications of the population considered may introduce model biases. Nevertheless the evolution of the client database of a banking group, for instance, is rarely taken into account by IRB models. This is why the consulting firm which participated to the workshop advocates for the adoption of a portfolio monitoring solution by banking institutions, wherein customer portfolios as well as credit and asset portfolios are regularly analysed to detect such structural changes.

Regarding the stability of the predictive model, a sub-project has been undertaken by the solution provider in order to provide KPIs as the basis of a monitoring and back-testing protocol of the hybrid models, in order to identify deviations of the model itself.

The stability of the hybrid model in fact only differs from that of the logistic regression model by the choice of the business rules embedded in it. That choice is made by the customer in interaction with the consulting firm, as both discuss the technical implications together. The customer may also choose during a model review to suppress a rule, for example to be more aligned with its risk appetite, or due to data quality problems identified on a variable involved in that rule.

Furthermore, the analysis of the model stability has shown that introducing business rules does not make the model less robust, provided those rules are guaranteed to apply only to the population segments identified. They also enable a specific monitoring of those population segments.

Lastly, initial studies suggest that a periodicity of 6 months for model updates would be adequate, both for credit scoring and for probability of default models.

Appropriate data management

In this kind of outsourced model-building solution, the validation of the resulting models, as well as that of the adequacy of data management, is ultimately the responsibility of the customer's compliance and risk departments.

There is thus no delegation of responsibility, nevertheless the regulatory requirements imposed to the end customer – particularly when it comes to explaining model predictions – are reported onto the solution provided by the third-party. Interestingly, the workshop participant has indicated that a project had been undertaken to set up an Ethics Committee involving large banking institutions among its customers, with the ultimate goal of producing an MRM (Model Risk Management) framework.

8.4.7. AI engineering methodology

The choice of an iterative model building workflow rather than a fully automated, single-step process is deliberate. Indeed, the solution designed by the workshop participant involves human intervention in the hybrid model optimization phase: this approach makes end-to-end automation of the build process impossible (while providing, according to the solution creators, benefits in terms of explainability and stability of the resulting model – see previous section).

Besides this lack of end-to-end automation, the engineering methodology relies on two foundations:

- On the one hand, the hybrid model building workflow follows a systematic approach and is developed according to industry standards.
- On the other hand, the tooling handed over to the customer will take the form of an information (model, data, and results) sharing platform, which enables the customer to be in the loop of all the decisions made and all the results obtained during model construction. The objective of this platform, still under construction at the time of this writing, is to provide an automated audit track of all exchanges with the customer.

The goal of this architectural choice is to make each stage of the model building traceable even after the model has been deployed in production, whether that stage has been automated or is performed by a human agent.

As for the risks induced by outsourcing (cf. section 5.4.2), they call for the following comments:

- The aforementioned model building method enables both reproducibility and auditability of the models produced.
- The quality of service is the customer's responsibility, as the customer ultimately decides to deploy the models and is in charge of their operational maintenance.
- Continuity of service and reversibility do not raise major difficulties either since the customer is able to revert to the regression model at any time, furthermore the evolution of business rules can be monitored independently from the model construction having been outsourced.
- Lastly, the risk of dependency towards the solution provider remains, specifically in its most fundamental aspect of technical knowledge: in this kind of situation, the end customer is responsible for developing and maintaining its expertise and know-how in order to control that risk.

8.5. Topic 3: Customer protection

This workshop was conducted with an insurance institution around a project pertaining to sales proposals: that project aims to produce prefilled quotes for home insurance.

8.5.1. Regulatory context

As mentioned in section 3.1.1, the duty to advise as defined by the IDD (Insurance Distribution Directive) imposes to sell an insurance product in accordance with the client's best interests. Therefore, the goal of technological innovation in that domain should be to make an offer consistent with the customer's needs and requirements – not to contribute to the creation of a demand.

8.5.2. Purpose of the exercise

The main challenge of this workshop was to shed light – by focusing on a specific use case – on the regulatory issues raised by the use of AI in the distribution of insurance products.

8.5.3. Objectives of the algorithm

For a customer who already subscribed to a contract, for example for an automobile insurance, the system implemented attempts to prefill a home insurance quote, including a "starting at" price.

8.5.4. Technical details

The specificity of this use case is its reliance on geographical data directly linked to real-estate sociology:

- Gridded data provided by INSEE (the French National Institute of Statistics and Economic Studies), including information such as the ratio of houses vs. apartments, the rate of home ownership, the average surface, the average household income (at the neighbourhood and commune levels).
- Data on buildings, acquired from a data provider, which gives the building surface and perimeter, from which a building shape is determined, and then a probability of house vs. apartment is estimated.
- The average number of rooms at the commune level.
- A postal address field, on which text analysis is performed in order to extract discriminating features for the house/apartment prediction.
- An email field, used for the same prediction.

The quote is prefilled with the following target variables, which are predicted iteratively (i.e. the 2nd one is predicted using the predicted value for the 1st variable, the 3rd using the first 2 predictions, and so on):

1. Home type: house or apartment
2. Customer status: owner or tenant
3. Number of rooms
4. Optional insurance of valuables
5. Year of construction

8.5.5. Validation process

Validation involved mainly the Compliance department, who performs consistency checks between the needs expressed by the customer on the one hand, and the risks declared in prefilled (then possibly amended) quote on the other hand.

8.5.6. Governance issues

The prefilled quote produced by the algorithm examined is leveraged by the insurance institution in several use cases:

- Sending via email a hyperlink to the quote.
- Processing incoming calls in order to perform portfolio cross-selling.
- Supporting outgoing telephone marketing campaigns.

The main governance issue is the respect of compliance requirements related to insurance product distribution, notably the duty to advise, which imposes that the motives for offering a particular product be exposed to the prospective customer, as well as the consistency between the product's characteristics and the customer's needs and requirements.

Particular attention should be focused on the human-machine interactions so that the subscription process based on prefilled information does not discourage the customer to express his or her needs[23], nor to verify the accuracy of the declared risks[24]. At the same time, regulation requires that any amendments (checking or unchecking of an option, or changes indicated by the customer) to the prefilled quote be faithfully reflected, as appropriate, in all other documents formalizing the gathering of customer needs and requirements and the insurance product offer.

These governance issues are illustrated by certain measures adopted during the system design and development. In order to ensure proper information of the customer, upon opening the prefilled quote a popup window explicitly enjoins the prospective customer to verify the information and to correct it if need be. The "insurance of valuables" option is particularly telling: it was initially checked in all quotes produced by the system, but it turned out that 60% of customers unchecked the box. It was therefore decided, again for the purpose of providing a quote as closely adjusted to the customer's intentions, to check or uncheck the box based on the model prediction on the "valuables" variable.

8.5.7. Evaluation methods and their implications

Explainability

Explaining an individual prediction to the customer does not represent, according to the insurance institution, a major issue in the case of predictive models intended for marketing: in the present case, rather than an explanation, an explicit validation request should be provided to the customer.

On the contrary, it seems important to provide an explanation for the model predictions – and more specifically its prediction errors – to the teams tasked with monitoring the system and with ensuring its compliance[25]. Indeed, a prediction error has a strong impact of the subscription process – a process which must be correctly understood by the customer: indeed, a failure to advice (or even liability) can be invoked when an erroneous prediction is not corrected by the insured and thus becomes a false declaration (albeit unintentional). Besides, prediction errors to the benefit of the insured generate – if they accumulate – an additional risk, this one of a financial and operational nature.

Performance

The predictive performance of the model is trivial to assess: it is measured as the classification accuracy according to each of the aforementioned target variables. By retaining the first three variables (with an error margin of one unit on the number of rooms), the model produces 90% of correct predictions.

Stability

This use case does not present any stability challenge, since input data are relatively static and the predictive power only has a minor business impact.

[23] In particular, an algorithm deemed efficient by the users may be endowed by them of a "prescriptive" power even though it has not been designed to that aim.

[24] At stake here are future potential disputes in case of a "false statement" whose origin would lie in the quote prefilling.

[25] However, the possibility to provide explanations to algorithmic decisions for purposes of internal control or external audit has not been explored during this workshop.

Appropriate data management

In terms of data management, the domain of insurance pricing defines forbidden variables. The absence of those variables from the resulting models should therefore be guaranteed, as well as the practical infeasibility of their inference from other predictor variables used by the model.

8.5.8. AI engineering methodology

The predictive models described here are deployed in production. Although they run in a production environment, this use of AI is not for an automated decision-making process: predictions are not provided continuously, instead a manual collection stage of the algorithm's output is necessary.

Thus, once the predictive model has been validated both functionally and technically, it is executed on a periodic basis and its results are used in the three situations previously described (email campaigns, incoming calls, and outgoing telephone campaigns).

9. Explainability vs. interpretability

The distinction between these two concepts is a frequent topic within the scientific literature, however there is no consensus on it.

Definition without a distinction

Burrell (2016) insists on the issue of the interpretability of algorithmic results, but without defining the terms in question. Doshi-Velez and Been Kim (2018) fail to distinguish the two terms as they define them in relation to each other. Nevertheless, their article strives to justify the necessity of categorizing various forms of interpretability. Similarly, Biran and Cottonn (2017) use a circular reasoning around both concepts: "*Explanation is closely related to the concept of interpretability: systems are interpretable if their operations can be understood by a human […]*".

While pointing out the lack of any formal definition, Bogroff and Guéguan (2016) define interpretability as the ability to explain or present stages using humanly understandable terms. For his part, Tim Miller (2018) offers a comprehensive analysis of both concepts. The introduction of the notion of degree allows to define interpretability as "*the degree to which an observer can understand the cause of a decision.*" Unfortunately, explainability is not defined according to the same notion, but as a way to obtain a human agent's understanding. Miller emphasizes the necessity for the reader to observe similarities and differences between the two concepts… but only after stating five lines prior that they would be used interchangeably.

Definition through distinction

Molnar (2019) lifts Miller's definition of interpretability so as to attempt distinguishing the two terms. He defines explainability as explanations of predictions being provided to individuals, and introduces the question of a "good explanation" in his book.

Bryce Goodman and Seth Flaxman, in *European Union regulations on algorithmic decision-making and a ''right to explanation''* (2017), implicitly distinguish the two concepts in their reading of GDPR's articles 13 to 15. They mention in particular that an algorithm operates by correlation and association, so that it performs predictions without providing any explanatory element of those correlations and associations. The difficulty that arises is thus that interpretation becomes difficult insofar as the algorithm works without having to explain its inner workings. The authors identify a tension between the right to access personal information collected (articles 13-15) and the right to collect data (article 22). Giving article 22 a disproportionate weight would lead to the development of a "black-box" society (Pasquale, 2015).

In his intervention at the *Institut de Recherche en Informatique de Toulouse* (2018), Laurent Serrurier links explainability to the technical characteristics of the algorithm, whereas interpretability is related to an ethical dimension. Explainability is thus a technical feature of the algorithm's complex nature, and interpretability refers to its social acceptability.

Likewise, in a talk given at the ACPR in 2019, Louis Abraham tackles Biran et Cottonn's definition which mixes both concepts *("Explanation is closely related to the concept of interpretability: systems are interpretable if their operations can be understood by a human, either through introspection of through* *a produced explanation."*), relates interpretability to the question "why" and explainability to the question "how".

Aurélien Garivier's 2018 article *"Toward a responsible artificial intelligence"* offers an explicit distinction when defining the two terms. As per the article *"A Berkeley View of Systems Challenges for AI"*, a decision rule is said to be interpretable if one can understand how it associates a response to observations; it is said to be explainable if one can understand on which elements the decision is grounded, possibly by using counterfactual reasoning.

Sub-distinction within interpretability

Lipton's 2017 article gives the most satisfying meaning to the concepts of interpretability and explainability. Rejecting the understanding of interpretability as a monolithic concept, Lipton introduces a continuum based on a number of logical criteria: trust in the algorithm's results, causality, transferability of knowledge, information contained in the decision, fairness of the decision. This framework enables to propose a concrete representation of the continuum between intelligibility and explanation.

10. Technical aspects of explainability

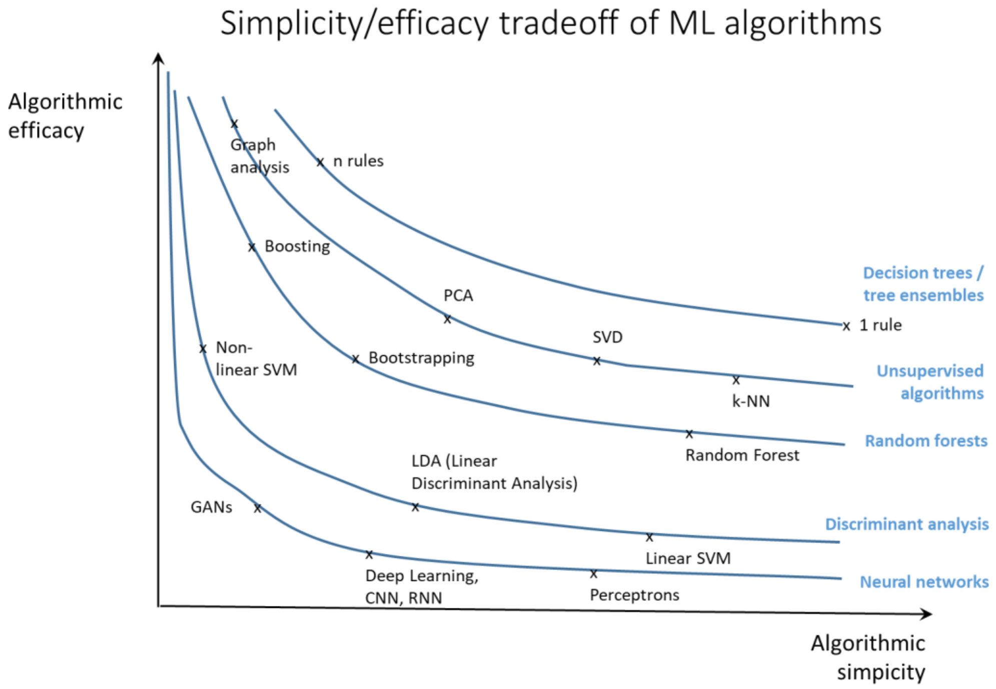
10.1. Trade-offs

This appendix describes the technical choices which arise after the appropriate level of explainability has been selected upon the introduction of AI in a business process. This description has a generic scope since the elements considered are not restricted to the financial sector. Two trade-offs are presented: between simplicity and efficacy of the ML algorithm on the one hand, between sobriety and fidelity of the chosen explanatory method on the other hand.

10.1.1. Simplicity/efficacy trade-off

A given type of ML algorithm may be more or less complex, in the sense of lending itself to an inspection of its inner workings. They can also vary in efficacy, measured as previously indicated using predictive performance or business performance metrics.

The following diagram attempts to illustrate the simplicity/efficacy trade-off among the most common ML algorithms:



Among the numerous simplifications and approximations operated by this diagram, the following points should be underlined.

Simplicity and efficacy metrics

On the one hand, ordering ML algorithm types in terms of their simplicity is highly subjective. Indeed, the size and structure of a model have a more significant impact on its explainability than the model type does, because understanding only part of the model is useless: thus a random forest comprising thousands of trees will typically be much more difficult to understand than a single-layer neural network composed of a dozen neurons.

On the other hand, the deterministic or stochastic nature of an algorithm is another essential criterion to take into account when assessing its efficacy. For instance, the results of a fundamentally stochastic algorithm depend on random sampling - not only for building training and evaluation datasets, but also within its procedure itself (for example, bootstrap methods contain re-sampling stages).

Lastly, it should be noted that the efficacy of a given algorithm type cannot be evaluated on a single-dimension scale either, since it depends on the use case considered (nature and volume of the data, choice of parameters, etc.)

Non-comprehensive taxonomy

It should also be emphasized that the representation of ML algorithms in the previous diagram does not pretend to be comprehensive. In particular, categories such as Reinforcement Learning have been excluded upfront because they are – to the best of our current knowledge – absent from solutions deployed as of today on the market.

On the other hand, unsupervised learning algorithms cannot be ignored. For example, graph analysis based on factoring company features allows to model the interdependency network generated by SMBs partaking in a P2P lending platform. The factoring technique used may be e.g. SVD (Singular Value Decomposition) or Latent Factor Model, and in any case that type of modelling demonstrated not only its descriptive value, but also its value as a predictor of credit default risk on such platforms (Ahelegbey, 2019). Furthermore, credit default risk is not easily amenable to traditional, non-ML-based model.

Decoupling between design and modelling

Lastly, the design of an algorithm can generally be decoupled from the structure of the resulting model: this is the strength and the innovation brought by hybrid models such as the one described in the workshop on probability of default (section 8.4).

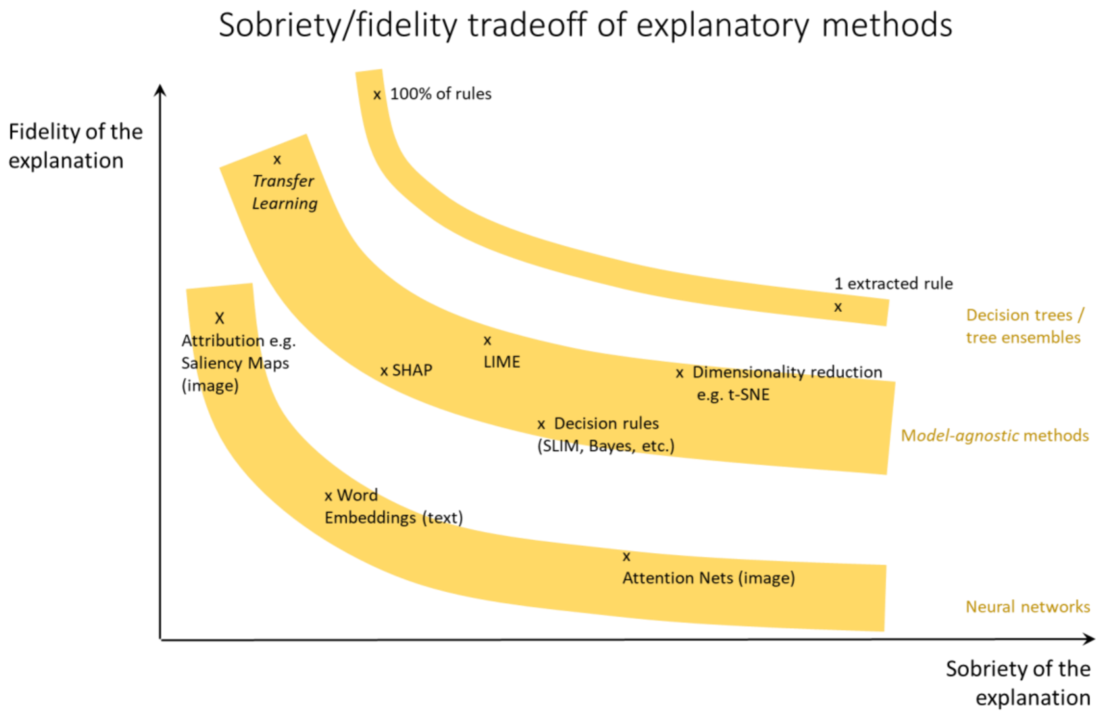
This approach consists of building a simple, intuitive model through iterative optimization by comparing it with a more efficient, often more complex model. The resulting model combines the best of both worlds, namely the performance (e.g. in terms of recall and precision) of a complex algorithm and the explainability (in terms of its interpretability and limited size) of the final predictive model.

10.1.2. Sobriety/fidelity trade-off

The explainability requirement induced by the introduction of ML into a business process is not limited to a simplicity/efficacy trade-off pertaining to the algorithm: the explanation should itself be intelligible and convincing to its intended recipients, suitable for the use case considered, and proportionate to the risk associated to the business process.

A trade-off is at play here as well. On the one hand, the explanation's fidelity (with respect to the algorithm which produced a given prediction) is imperfect since the algorithm's behaviour is necessarily simplified when its output is explained in terms of certain characteristics of the individual or transaction considered. On the other hand, the sobriety of the explanation, that is its intuitiveness and intelligibility by a layperson, is both subjective and constrained in practice.

The following diagram attempts to represent the sobriety/fidelity trade-off for an explanation according to the type of ML algorithm and the type of explanatory method. A few "corridors" are drawn to show that for a given algorithm type, some explanatory methods will deviate slightly from the general trends.



10.2. Reaching a high explanation level

10.2.1. Feasibility of replication

It should be noted that explanation level 4 (replication) aims to identically reproduce the model's behaviour, and not to understand its inner workings in their fullest detail – which may prove impossible for certain models, typically deep neural networks.

Interestingly, some of the financial actors met during the exploratory works led by the ACPR implemented a replication method as early as the design and initial validation phases of their algorithms, hence upstream from the internal control or audit procedures. More specifically, they opted for implementing their ML algorithms in multiple (in some cases three) languages, which is a software engineering technique classically used for particularly critical components of a system.

10.2.2. The problem of software dependencies

Besides, a problem arises whenever a code review is in order (i.e. for levels 3 and 4). This problem is not specific to AI algorithms and comes up in virtually any well-conceived software audit mission: multiple external software libraries, tools and components are invoked by the code analysed, and their review ranges from difficult (in the case of open source software) to impossible (in the case of closed-source code). Even a simple logistic or linear regression algorithm uses several third-party libraries, and the problem is amplified for sophisticated algorithms, which incidentally also require a higher explanation level.

In conclusion, reaching a level-3 or level-4 explanation is challenging in most situations, and the challenge is made more difficult under certain circumstances: when the algorithm relies on third-party libraries or products, and when the audit mission needs to cover the entire model building workflow and not just the resulting model.

An approach sometimes mentioned to facilitate this kind of in-depth analysis consists of setting up a certification process of off-the-shelf ML components, similarly e.g. to how software security components must be tried-and-tested and officially approved prior to being embedded into critical applications. At any rate, the detailed code analysis suggested for level-4 explanations (replication) should also focus on the use of such off-the-shelf libraries, with a crucial point being the hyper-parameter optimization stage insofar as it significantly impacts the algorithm's accuracy.

11. Review of explanatory methods in AI

This review does not pretend to be comprehensive: it is limited to the use of AI in the financial sector, besides it aims to paint a picture of the use cases which are deployed in practice – whether currently in production or simply at an experimental stage.

The frame of reference here is an ML algorithm. Explanatory algorithms can be classically grouped into three categories according to where they come up within the model's lifecycle:

1. Pre-modelling explanatory methods aim to describe the data used when building the models.
2. Explainable modelling contributes to the production of intrinsically more explainable models.
3. Lastly, post-modelling explanatory methods attempt to yield satisfactory explanations from previously built and trained models.

11.1. Pre-modelling explanatory methods

Upstream from the learning stage of an ML model, a somewhat limited form of explainability can be provided, whose goal is to illustrate the data used by the algorithm. The most common methods used for this purpose are the following.

Exploratory data analysis

Exploratory data analysis often relies on data visualization, which enables to reveal characteristics of the data – even when they are potentially hidden from descriptive statistics.

These methods are both model- and domain-agnostic. Within the financial sector, they are particularly useful for detecting and mitigating undesired biases (cf. section 3.1.2). Potential sources of such problematic biases are numerous (Kamishima, 2012): direct, indirect or latent dependency on sensitive variables, sampling biases (the most difficult case to detect) or labelling biases in the training data, or imperfect convergence of the training stage.

Dataset documentation

Several dataset documentation standards have been suggested, either applicable to AI models (Mitchell, 2019) or to associated services (Hind, 2018).

This kind of approach, based on a thorough, concise and formal documentation of the datasets and services, is suitable for level-1 explanations as described in this document (cf. section 3.4.4).

However, the technical focus of the standards proposed thus far makes them ill-suited for the customers and end users of AI algorithms: instead, they are intended for the creators of those tools, or even for the individuals in charge of monitoring their operational behaviour. This standardization effort is nevertheless recent and likely to evolve in a near future.

Dataset summarization methods

In order to facilitate the mental representation and the interpretation of datasets, particularly the voluminous and heterogeneous ones, certain dataset summarization methods may be used, ideally as a complement to the aforementioned exploratory analysis and documentation methods.

Examples of dataset summarization methods are:

- For textual data, automatic document summarization and classification.

- For images, visual scene synthesis.
- For any data type, the extraction of representative (or typical) exemplars from a dataset, and of particularly atypical exemplars as well (respectively called prototypes and criticisms: Kim, 2016).

Explainable feature engineering

This last type of pre-modelling explanatory method stems from the observation that an explanation for a predictive model is only as good as the predictive features it relies on. Therefore, particular care must be taken to the feature engineering stage when designing an ML system, i.e. to the construction of predictor variables from original variables in order to adequately re-structure the training data for the algorithm.

Two such methods should be mentioned (Murdoch, 2019):

- The intervention of domain experts, who are sufficiently knowledgeable about the source data to extract variables (combination of other variables, intermediate computation results, etc.) which increase a model's predictive accuracy while maintaining the interpretability of its results. In other words, human expertise enables in certain cases to sidestep the usually inevitable trade-off between efficacy and explainability of an ML model (cf. section 10.1.1).
- A modelling-based, automated approach: usual data analysis techniques are then used, such as dimensionality reduction and clustering, so as to extract predictor variables as compact and representative as possible.

11.2. Explainable modelling

Some methods enable simultaneously training the predictive model and building an associated explanatory model. This category of explanatory method is referred to as explainable modelling.

Such methods are however far less frequently implemented than pre- and even more post-modelling explanatory approaches, for several reasons:

- Explainable modelling requires access to the source code which produces the predictive model, and the possibility to modify the algorithm. On the contrary, access to the model itself is sufficient for post-modelling explanatory methods, which makes them much more widely applicable.
- Explainable modelling is useful when explanations are necessary as early as the design phase of the ML algorithm, which demands a more mature engineering methodology and adequate planning during the introduction of AI into a business process.
- Lastly, explainable modelling is not very suitable for audit, all the more so when the predictive model is only available as a black box, without a documentation of the algorithm itself.

The primary, highly ambitious goal of explainable modelling is to avoid as much as possible the already mentioned trade-off between efficacy and explainability, as they strive to provide additional explainability without necessarily sacrificing predictive accuracy.

A few methods for explainable modelling are described in what follows.

Intrinsically explainable models

An intrinsically explainable model can be chosen from the outset, for example linear models or decision-tree-based models. This is the most trivial kind of explainable modelling approach, assuming that the simplicity/efficacy trade-off is kept in mind, and that the specific model produced by the algorithm is actually explainable. The latter point is not always guaranteed: in some cases, adopting an explainable family of models is not sufficient since it may lead to a model with too many dimensions to remain intelligible.

Hybrid explainable models

Hybrid explainable models are only applicable to a specific model type, namely neural networks. The following types of models belong to this category:

- Deep k-NN (Papernot and McDaniel, 2018) extracts the internal representation of a neural network within each of its layers in order to illustrate how the final result is obtained (in the last layer). A variant of this approach is the Deep Weighted Averaging Classifier.
- SENN (Self-Explaining Neural Networks: Alvarez-Melis, 2018) uses neural networks to simultaneously train the predictor variables, the weights, and the aggregation method of a linear model. A variant is the Contextual Explanation Network (Al-Shedivat, 2018).

Joint prediction and explanation

This approach consists of training the model to produce both a prediction and an explanation for this prediction. It has recently received considerable attention, despite two major limitations. Firstly, not only does the ML algorithm need to be modified, but explanations must be provided for the entire training dataset, which is often unrealistic. Secondly, the explanations produced are only as accurate and relevant as the information provided by human agents for training the hybrid model, they do not necessarily justify the genuine, internal workings of the predictive model.

The following methods fall under this joint prediction/explanation approach:

- TED (Teaching Explanations for Decisions: Hind, 2019) associated to each training data point the motive behind the resulting prediction. A variant is the generation of multimodal explanations (Park, 2018).
- Data-type-specific methods: these include Visual Explanations (Hendricks, 2016) for object recognition in images, or the generation of concise explanations in natural language (e.g. English) for a predictive model using textual source data (Lei, 2016).

Architectural adjustment methods

Methods relying on architectural adjustments are mostly specific to Deep Learning (which is as of today relatively infrequent in the financial sector).

A few of them are nonetheless worth mentioning, such as Attention-Based Models which aim to identify the most important feature groups within input data, be they images, textual data, or – more relevantly in the financial sector – time series. Some studies (Jain, 2019) however illustrate the limits of this approach in terms of performance of the resulting model.

Regularization methods

Regularization methods are typically used to enhance the performance of an ML model, however some kinds of regularization enable to improve model explainability.

For example, the decision boundary of a model may be constrained during training to be approachable by a decision tree, which makes future predictions easily comprehensible by a human (Wu, 2017). Another example are methods which orient model training to assign more weight to the predictor variables labelled as most important by a domain expert (Ross, 2017).

Training/modelling decoupling

Of particular note are specialized approaches which decouple the training stage of an ML algorithm from the structure of the resulting model.

An example of such approach is the hybrid method described in the workshop on probability of default (section 8.4). An advanced model, with low explainability by nature, is trained to achieve high predictive accuracy, after which domain experts extract a number of business rules to augment an intrinsically explainable model (e.g. a decision tree) with a number of "overrides". The resulting system thus benefits both from the accuracy of a complex algorithm and from the explainability of a simple predictive model.

11.3. Post-modelling explanatory methods

Methods operating on previously-trained ML models are *de facto* the most commonly intended meaning for explanatory methods in general. Their goal is to provide *post-hoc* explanation which justify or illustrate a given result (or a set of results) produced by an ML model. The model is thus considered as the object studied, on which changes can not be made (contrary to explainable modelling approaches from section 11.2) and whose data can not be manipulated (contrary to pre-modelling approaches from section 11.1).

Two main criteria are used to distinguish post-modelling explanatory methods. Firstly, their local or global scope:

- Local explanatory methods provide an explanation for a decision made on a particular input data point (for instance, why a given credit application was granted to the applicant).
- Global explanatory methods attempt to simultaneously explain the entirety of possible decisions (in this case, what are the general characteristics of the respective outcomes – acceptance or denial – of credit applications).

The second criterion is whether a method is applicable to any type of ML model (model-agnostic methods) or only to specific type of model or algorithm (model-specific methods).

11.3.1. Local explanatory methods

Black-box methods

Black-box methods, also called model-agnostic, are applicable to any type of model. They may consist of a simple classifier (for example a Bayesian classifier trained on Parzen windows), or be more sophisticated (a number of them operate by perturbing the model then observing the influence of predictor variables).

The following techniques are among the most common model-agnostic explanatory methods:

- Naive Bayes Models, which are often crude in comparison to the next ones.
- LIME (Locally Interpretable Model-Agnostic Explanations) works by constructing an intermediate representation domain between the ML model and the "real-world" model so as to find the optimal trade-off between fidelity of the model explanations and simplicity of the explanations (whose purpose is to be intelligible by domain experts who are not necessarily technically savvy).

- SHAP combines game theory (Shapley values) with the optimization of credit allocation in order to explain the influence of each predictor variable on the predicted values, also in a model-agnostic manner (Lundberg, 2017).
- Variants of the SHAP method, for example adapted to data structured as a network (Chen, 2019).
- Causal interpretation methods, which compute the marginal influence of each predictor variable and the joint influence of variable pairs (Datta, 2016).
- SLIM (Supersparse Linear Integer Models) which selects decision rules so as to optimize the accuracy of a binary classifier under constraints on the number of variables and their relative weights.

It should be noted that even the most commonly used local explanatory methods, such as LIME and SHAP which are both based on model perturbations, encounter practical limitations in terms of security (Dylan, 2020). In particular, they are vulnerable to adversarial attacks (cf. appendix 12) which can produce models including discriminatory biases on which the explanations generated are reassuring or even indistinguishable from the explanations produced on an unbiased model.

Some black-box explanatory methods are also specific to models operating on NLP, and generally provide either numeric explanations or explanations in the form of a textual example:

- An adaptation of the LIME method to NLP (Ribeiro, 2016) provides explanations as the degree of importance of each predictor variable.
- A generative method (Liu, 2018) provides explanations as a simple textual example.

Model-specific methods

A number of local explanatory methods are specific to a type of ML model.

It should first be noted that some models are directly interpretable:

- Logistic regressions.
- Linear regressions and variants such as GLM (Generalized Linear Models), provided their density is limited.
- Additive models such as GAM (Generalized Additive Models).
- Decision trees and random forests, at least when they are limited in depth and volume.

A number of explanatory methods are specific to Deep Learning models:

- Explanations in the form of surrogate models, particularly decision trees which approximate the neural network (Craven, 1995).
- Explanations based on attention mechanisms (Choi, 2016).
- Explanations which attribute decisions of the neural network to certain predictor variables (Shrikumar, 2017).

Lastly, the following methods are domain-specific:

- Explanations for NLP algorithms based on Recurrent Neural Networks (Strobelt, 2018).
- Explanations for CV algorithms, for example Interpretable Units (Bau, 2017), Uncertainty Maps (Kendall, 2017), or Saliency Maps (Adebayo, 2018).

Counterfactual explanations

Counterfactual explanations have their own place among methods aiming to explain an ML algorithm, insofar as they are the only ones involving causal relations[26] (and not just explanations grounded in statistics or inferences generalized from large data volumes).

More precisely, a counterfactual explanation to prediction Y, generated by a model from input data `X`, is given by input data `X'` as close as possible to `X` which would have resulted in prediction `Y'` different from Y. In general, `Y` is an unfavourable outcome (prediction or decision), for example a low credit score computed from `X` resulting in the credit application being denied. A relevant explanation (to the creator of the system, to an auditor, but above all to an individual impacted by the outcome, in this case the applicant who requested the credit) should then answer the question: what change as minimal as possible in the credit application would have led to its acceptance? Thus, rather than a local explanation which quantifies the influence of various predictor variables (age, income, credit history, etc.) on the negative outcome, a far more useful, practical and simple explanation is obtained, for example "*if the household income had been this much instead of that much, the credit would have been granted.*"

Certain methods for generating counterfactual explanations even goes beyond this definition (McGrath, 2018):

- Some methods produce positive counterfactual explanations, i.e. which apply in cases where the original decision `Y` is favourable to the individual considered. Using the previous example, `Y'` corresponds to the credit application being denied, thus the counterfactual explanation indicates a safety margin for the favourable outcome. This kind of explanation may be useful to make an informed decision e.g. to request another credit in the future given that the initial application has been accepted.
- Another enhancement is achieved by weighing explanatory factors based on their variability. Using yet again the credit scoring example, if the individual has proven to be better able to reduce their personal expenditures than to increase their revenue, then this enhancement method would produce an explanation such as *"if monthly expenses had been cut by half, then the credit would have been granted."* This specific explanation is indeed more useful than one involving other features such as the household income, in that it is more directly actionable.

Ideally, counterfactual explanatory methods should be applicable to algorithms studies as black boxes. Certain methods do in fact satisfy this condition under well-determined situations (Wachter, 2018).

11.3.2. Global explanatory methods

Global explanatory methods provide an explanation to the entirety of decisions made by an ML model: for example, what is the contribution of the "age" variable to the decisions to accept or deny credit applications over the set of all applications.

Global explanatory methods may be useful to internal control teams or to an auditor in order to obtain an understanding of the general behaviour of an algorithm, however they usually show their

[26] This ability to tackle causality is very promising for the deployment of AI in general, and within the financial sector in particular. For example, the explainability of internal risk models implemented by banking institutions would be reinforced if those models enabled to assess causal relations. Causal inference is *de facto* at the core of the concerns in empirical economy since at least 25 years. Nevertheless it is missing from commonly-used AI models, as well as from the more classical models currently deployed by banks.

limitations when compared to the study of a single, concrete use case (using a local explanation) or several concrete use cases (for example to compare the algorithmic results on two individuals and detect a potential inequality of treatment).

Global explanatory methods are also very difficult to materialize in practice. Such methods exist for specific model types: for example it is possible to extract from deep neural networks a set of decision rules which is easy to interpret and, according to the situation, relatively faithful to the Deep Learning model considered (DeepRED : Zilke, 2016).

In addition, few methods are able to provide a global explanation independently from the type of the model being studied. This is however the case of Partial Dependence Plots (PDP), which show the marginal effect of a given variable on the model predictions (Friedman, 2001).

12. Review of potential attacks against an ML model

ML security if a very recent field of study, but important enough to have been the object of a taxonomy. This taxonomy is nevertheless in constant evolution, given the changing nature of the field (Papernot, 2018). Among the most notorious attacks against ML models, the following categories can be distinguished (with an example scenario is given for each type):

- Causative attacks (a.k.a. Data Poisoning): training data are altered (modified feature values, new features created, etc.)
    - Causative integrity attacks are a subcategory of causative attacks. They are used e.g. to grant generous loans or low insurance premiums to malicious individuals.
    - Causative availability attacks are another subcategory of causative attacks. They are used e.g. to discriminate against a population group by denying them the same benefits as the rest of the population.
- Watermark Attacks: a malicious actor modifies the code of the ML algorithm.
    - Watermark integrity attacks: for example, conditions are introduced on certain input data features so as to trigger advantageous outcomes in chosen cases.
    - Watermark availability attacks: for example, rules are injected into the code in order to suppress favourable outcomes for the target population group.
- Surrogate Models Attacks
    - Inversion attacks are the equivalent of reverse engineering for ML model.
    - Membership tests are another type of surrogate model attacks which operate by inferring whether input data points belong to the original training set.
- Adversarial attacks construct synthetic examples in order to avoid a detrimental outcome – or alternatively to obtain a favourable outcome.
- Impersonation Attacks work by injecting data corresponding to a real identity (or a composite of several real identities) in order to usurp that identity.

A particularly interesting aspect of ML security is that defences against them tend to have a beneficial side effect (Hall, 2019), namely that they bring the ML model closer to satisfying the four design and development principles mentioned in this document (section 3): appropriate data management, performance, stability, and explainability.

To give but one example, a defence against data poisoning attacks is the RONI method (Reject On Negative Impact). It works by rejecting training data which decreases the model accuracy (Barreno, 2010), hence it also protects against degrading the model performance due to training data drift. As an illustration, a facial recognition algorithm secured by RONI will exclude from its training set a series of pictures, each associated to an ID document, which would significantly lower its precision: this contributes to ensuring the integrity but also the performance of the model, which could for example be used by a banking institution to remotely identify new customers (a use case commonly known as "KYC at a distance").

Bibliography

Louis Abraham. *In Algorithms We Trust*. ACPR (21 mars 2019).

Peter Addo, Dominique Guégan, Bertrand Hassani. *Credit Risk Analysis using Machine and Deep Learning models*. ffhalshs-01719983f (2018).

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: *Sanity Checks for Saliency Maps*. NeurIPS 2018: 9525-9536 (2018).

AEAPP. *Final Report on public consultation No. 19/270 on Guidelines on outsourcing to cloud service providers*. EIOPA-BoS-20-002 (2020).

Daniel Felix Ahelegbey, Paolo Giudici, Branka Hadji-Misheva. *Latent Factor Models for Credit Scoring in P2P Systems*. Physica A: Statistical Mechanics and its Applications No. 522 (10 February 2019): pp. 112-121 (2018).

Maruan Al-Shedivat, Avinava Dubey, Eric P. Xing. *Contextual Explanation Networks.* arXiv:1705.10301v3 [cs.LG] (2018).

David Alvarez-Melis, Tommi S. Jaakkola. *Towards Robust Interpretability with Self-Explaining Neural Networks.* arXiv:1806.07538v2 [cs.LG] (2018).

0 David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, Klaus-Robert Muller. *How to Explain Individual Classification Decisions.* J. Mach. Learn. Res. 11: 1803-1831 (2009).

Marco Barreno, Blaine Nelson, Anthony D. Joseph, J.D. Tygar. *The security of machine learning.* Mach Learn (2010) 81: 121–148 DOI 10.1007/s10994-010-5188-5 (2010).

Robert P. Bartlett, Adair Morse, Richard Stanton, Nancy Wallace. *Consumer-lending discrimination in the FinTech era* (No. w25943). National Bureau of Economic Research (2019).

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba. *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. CVPR 2017: 3319-3327 (2017).

Roland Berger. *The road to AI Investment dynamics in the European ecosystem*. AI Global Index (2019).

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh Ruchir Puri, José M. F. Moura, Peter Eckersley. *Explainable Machine Learning in Deployment.* arXiv:1909.06342 [cs.LG] (2019)

Or Biran, Courtenay V. Cotton. *Explanation and Justification in Machine Learning: A Survey* (2017).

Alexis Bogroff, Dominique Guégan. *Artificial Intelligence, Data, Ethics: An holistic Approach for Risks and Regulation*, HAL (2019)

Jenna Burrell. *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data & Society (2016).

Cambridge Judge Business School. *The Global RegTech Industry Benchmark Report* (2019).

Cambridge Judge Business School, World Economic Forum. *Transforming Paradigms A Global AI in Financial Services Survey* (2019).

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan. *L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data.* ICLR (2019).

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart. *RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism.* NIPS 2016: 3504-3512 (2016).

Mark Craven, Jude W. Shavlik. *Extracting Tree-Structured Representations of Trained Networks*. NIPS 1995: 24-30 (1995).

Wei Dai, Isaac Wardlaw. *Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking.* Information Technology, New Generations. Advances in Intelligent Systems and Computing. 448. pp. 439–450. ISBN 978-3-319-32466-1 (2016).

Anupam Datta, Shayak Sen, Yair Zick. *Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems*. In Security and Privacy (SP), 2016 IEEE Symposium on, pp. 598–617. IEEE (2016).

Doshi-Velez, Been Kim. *Towards a Rigorous Science of Interpretable Machine Learning* (2017).

EBA. *Draft recommendations on outsourcing to cloud service providers under Article 16 of Regulation (EU) No 1093/20101*. EBA/CP/2017/06 (2017).

European Commission High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy Artificial Intelligence* (2019).

Jerome H. Friedman. *Greedy function approximation: A gradient boosting machine.* Annals of statistics (2001).

Donna Fuscaldo. *ZestFinance Using AI To Bring Fairness To Mortgage Lending* (2019).

Aurélien Garivier. *Toward a responsible artificial intelligence.* Institut Mathématique de Toulouse (26 mars 2018).

Bryce Goodman, Seth Flaxman. *European Union regulations on algorithmic decision-making and a ''right to explanation''* (2017).

Dominique Guégan, Bertrand Hassani. *Regulatory Learning: how to supervise machine learning models? An application to credit scoring*. ffhalshs-01592168v2f (2017).

Patrick Hall. *Proposals for model vulnerability and security.* O'Reilly (2019).

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell. *Generating Visual Explanations.* arXiv:1603.08507v1 [cs.CV] (2016).

Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, Kush R. Varshney. *Increasing Trust in AI Services through Supplier's Declarations of Conformity.* CoRR abs/1808.07261 (2018).

Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, Kush R. Varshney. *TED: Teaching AI to Explain its Decisions.* arXiv:1811.04896v2 [cs.AI] (2019).

Sarthak Jain, Byron C. Wallace. *Attention is not Explanation.* arXiv:1902.10186v3 [cs.CL] (2019).

Konstantinos Koutroumbas, Sergios Theodoridis. *Pattern Recognition (4th ed.).* Burlington. ISBN 978-1-59749-272-0 (2008).

C. Jung, H. Mueller, S. Pedemonte, S. Plances, O. Thew. *Machine learning in UK financial services*, Bank of England & Financial Conduct Authority (2019).

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, Jun Sakuma. *Fairness-Aware Classifier with Prejudice Remover Regularizer.* P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 35–50 (2012).

Alex Kendall, Yarin Gal. *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?* NIPS 2017: 5580-5590 (2017).

Faye Kilburn. *BlackRock to use machine learning to gauge liquidity risk* (2017).

Been Kim, Rajiv Khanna, Oluwasanmi Koyejo. *Examples are not Enough, Learn to Criticize! Criticism for Interpretability.* 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain (2016).

KPMG. *AI Compliance in Control* (2019).

Tao Lei, Regina Barzilay, Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP (2016).

Zachary C. Lipton. *The Mythos of Model Interpretability* (2017).

Hui Liu, Qingyu Yin, William Yang Wang. *Towards Explainable NLP: A Generative Explanation Framework for Text Classification*. CoRR abs/1811.00196 (2018).

Lloyd's, *Taking control - Artificial intelligence and insurance*. Emerging Risk Report (2019).

Scott M. Lundberg, Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. NIPS 2017: 4768-4777 (2017).

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, Su-In Lee. *Explainable AI for Trees: From Local Explanations to Global Understanding.* arXiv:1905.04610v1 [cs.LG] (2019).

MAS (Monetary Authority of Singapore). *Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector.* (2019).

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamia, Zhao Shen, Freddy Lécué. *Interpretable Credit Application Predictions With Counterfactual Explanations.* arXiv:1811.05245v2 [cs.AI] (2018).

Tim Miller. *Explanation in Artificial Intelligence: Insights from the Social Sciences* (2018).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, BenHutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. *Model Cards for Model Reporting.* In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (2019).

Christoph Molnar, *Interpretable Machine Learning — A Guide for Making Black Box Models Explainable* (2019).

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yua. *Interpretable machine learning: definitions, methods, and applications.* arXiv:1901.04592v1 [stat.ML] (2019).

Nicolas Papernot, Patrick McDaniel. *Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning.* arXiv:1803.04765v1 [cs.LG] (2018).

Nicolas Papernot. *A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private.* Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (2018).

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, Marcus Rohrbach. *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence.* arXiv:1802.08129v1 [cs.AI] (2018).

Keyur Patel, Marshall Lincoln. *It's not magic: Weighing the risks of AI in financial services*, Centre for the Study of Financial Innovation (2019).

James Proudman. *Cyborg supervision-the application of advanced analytics in prudential supervision*, Bank of England (2018).

PwC. *Opportunities await: How InsurTech is reshaping insurance.* Global FinTech Survey (2016).

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. *Model-agnostic interpretability of machine learning.* ICML Workshop on Human Interpretability in Machine Learning (2016).

Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Explainable NLP KDD 2016: 1135-1144 (2016).

Andrew Ross, Michael C. Hughes, Finale Doshi-Velez. *Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations.* arXiv:1703.03717v2 [cs.LG] (2017).

Lukas Ryll, Sebastian Seidens. *Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey*. arXiv:1906.07786 (2019).

Laurent Serrurier. *Un point sur l'explicabilité et l'interprétabilité en (DEEP…) Machine Learning*, IRIT (12 novembre 2018).

Blake Shaw, Tony Jebara. *Structure Preserving Embedding.* Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada (2009).

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov. *Membership Inference Attacks Against Machine Learning Models.* 2017 IEEE Symposium on Security and Privacy (2017).

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje. *Learning Important Features Through Propagating Activation Differences*. ICML 2017: 3145-3153 (2017).

Justin Sirignano, Apaar Sadwhani, Kay Giesecke. *Deep learning for mortgage risk* (2018).

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju. *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.* arXiv:1911.02508 [cs.LG] (2020).

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, Alexander M. Rush. *LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks.* IEEE Trans. Vis. Comput. Graph. 24(1): 667-676 (2018).

Erik Štrumbelj, Igor Kononenko. *An efficient explanation of individual classifications using game theory*. Journal of Machine Learning Research, 11:1–18 (2010).

Tapestry Networks. *Why banks can't delay upgrading core legacy banking platforms* (2019).

Berk Ustun, Cynthia Rudin. *Supersparse Linear Integer Models for Optimized Medical Scoring Systems*. Machine Learning 102.3: 349–391 (2015).

Sandra Wachter, Brent Mittelstadt, Chris Russell. *Counterfactual explanations without opening the black box : automated decisions and the GDPR.* Harvard Journal of Law & Technology Volume 31, Number 2 Spring 2018 (2018).

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual*

Attention. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057 (2015).

Jan Ruben Zilke, Eneldo Loza Mencía, Frederik Janssen. *DeepRED – Rule Extraction from Deep Neural Networks.* In: Calders T., Ceci M., Malerba D. (Eds) Discovery Science. DS 2016. Lecture Notes in Computer Science, vol. 9956. Springer, Cham (2016).

Thanks

The authors of this document wish to thank the participants to the exploratory works described herein (domain experts, data scientists, validation and compliance teams, as well as their managers) for responding to the call for applications and for their active contribution to the discussions, presentations and workshops.

The authors also wish to thank their ACPR colleagues who contributed to the brainstorming and analyses which resulted in this document, and specifically: Jean-Philippe Barjon, Emmanuelle Boucher, Nicolas Carta, Laurent Clerc (AML-CFT Unit), Richard Diani, Thierry Frigout, Cyril Gruffat, Gauthier Jacquemin, Boris Jaros, Matthias Laporte, Farid Oukaci, Jérôme Schmidt, and Pierre Walckenaer.