The Challenge of AI Model Validation: Our Approach

AMSTERDAM DATA COLLECTIVE

Abstract

AI is taking an increasingly central role in the strategy of financial institutions. Most institutions have data science teams devising innovative ways to employ Artificial Intelligence (AI) in meaningful ways. However, only a small fraction of the modelling efforts ends up being implemented and yield meaningful results. Our advice is to critically assess existing model governance frameworks and revise them to facilitate AI modelling. In this article we focus on the validation process. The table below summarizes some challenges that arise when validating an AI model compared to traditional models and how Amsterdam Data Collective would approach this.

Validation dimension	Traditional models	AI models	Our approach
Model input	• What goes into the model and does the dataset cover the expectations? • What is the quality of the data? • Is the data complete, accurate and representative? • How do the modellers deal with outliers and missing values?	• Datasets with both higher volume and dimensionality • Unstructured data • Modern data pipeline • Unintended bias might arise	• Focus on data quality strategy • Assess input data on potential bias • Understand and test data pipeline • Assess unstructured data using AI techniques
Model soundness	• What is the model structure and is this mathematically correct? • What assumptions have been made and do they hold?	• Black box: model can easily become complex and hard to understand. • Modelling techniques that are not as widely understood as traditional models. • Harder to boil down to assumptions.	• Use ‘standard’ components • High standard of coding • Clear documentation, including components like a proper Git flow. • Explainable AI
Model output	• Test calibration accuracy • Discriminatory power, • Robustness of predictions • Stability of model estimates	• Link between output and input less straightforward • Other metrics to assess outcomes • More extensive monitoring • Prone to overfitting	• Automatic triggers to test stability continuously • Rely on more computationally intensive methods • K-fold cross validation • Benchmark accuracy and discriminatory power with more interpretable models
Regulatory compliance	• Is the model in line with the most recent external and internal regulation? • Ideally a regulatory checklist has been used and can be assessed.	• More freedom in modelling choices • Authorities promote use of AI if added value is clear • Guidelines sometimes available but less clear	• Use and adhere to available guidelines for use case • Assess internal guidelines • Iterate often with development team and start early
Model implementation	• Is the model implemented correctly and in line with model design? • Is the model running in a suitable IT environment?	• Design flaws prevent models going into production. • Continuous monitoring • More flexible and versatile IT platforms.	• Review the set requirements • Multidisciplinary validation team with data engineering capabilities • Increased focus on monitoring and change management.

Introduction

Artificial Intelligence (AI) has been irrevocably adopted by the financial industry. Not only every Fintech, but every bank, trader or regulator has data science teams devising innovative ways to employ AI in their daily business. And rightfully so: With the ever-increasing data availability, more and more use cases can benefit from AI techniques, as visualized in figure 1. However, this adoption of AI comes with a challenge. Most modelling efforts do not make it to implementation and do not yield the intended results. An important step in employing AI effectively, is to incorporate AI models in the, often more traditional, model governance frameworks.

We find that organizations with model governance frameworks that are not tailored to AI models, are rarely successful at implementing and scaling their AI efforts. Traditionally, model governance frameworks implemented in banks focus on financial risk models, which typically are based on econometric methods . Even though the similarities between econometric- and AI methods are numerous, some key differences make the current model governance frameworks, and with it the validation process, not suitable for AI models. Some of these key differences are:

- The Black-box characteristic of AI models are incompatible with requirements on interpretability.
- The increase in complexity of both algorithms and data mean that the model is highly dependent on modelling choices, even more than for traditional models.
- Regulatory guidance on AI models is not well defined yet. Appropriate governance frameworks can compensate for this.

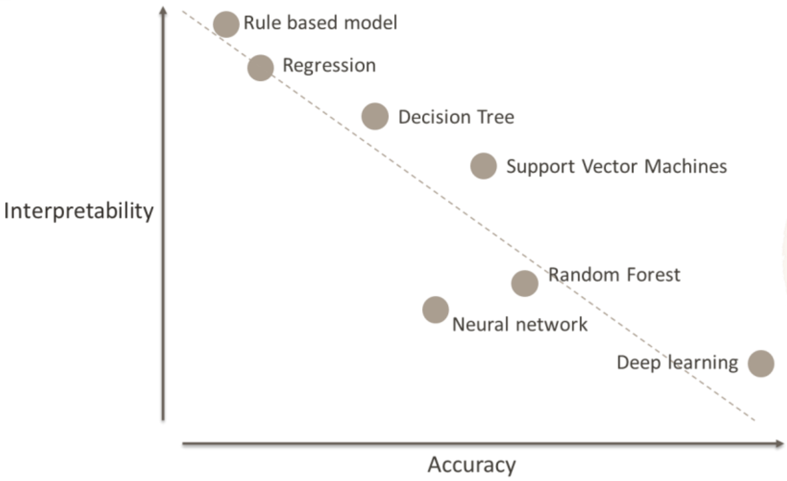

Figure 1: When does it make sense to apply AI to a use case despite the added complexity in model governance? With the ever-increasing amount of data, more and more use cases might benefit. For a given use case, the trade-off between interpretability and accuracy should be assessed. If the technique is above the dashed line, the trade-off might be worth it. However, the trade-off in interpretability and overall complexity raises a challenge for model validation.

AI is an umbrella term. In this article, AI specifically refers to modelling techniques that are employed when the data availability is high and the patterns in the data are not straightforward and easily captured.

Instead of fitting AI solutions into frameworks designed for simpler models, the framework should be redesigned to accommodate more complex techniques.

As Figure 2 shows the validation process is a central aspect of the model management cycle. This article aims to shine some light on the general challenges that come with embedding AI models in existing validation frameworks.

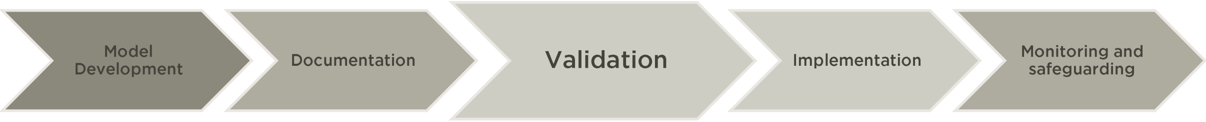

Figure 2: General steps in a model governance framework.

USE OF AI MODELS TO DETECT AML AND CFT

An interesting example of an area in which the application of AI models is gaining a lot of traction is Anti-Money Laundering/Combating the Financing of Terrorism (AML/CFT). Banks in Europe have been under increasing scrutiny from regulators on this subject. Financial institutions are encouraged by regulators to employ AI to detect potential money laundering or financing of terrorism more effectively (DNB, 2017), and, in general guidance on AI is increasing, see for example the EBA guideline on advanced analytics (EBA, 2020). However, requirements such as transparency and explainability are still high on the agenda. This makes AML/CFT a field in which model management, regulation and the need for AI come together in an unprecedented way. Throughout this article this exemplary field will be used to clarify and illustrate the subject.

FIVE STEPS WITHIN VALIDATION PROJECTS

Amsterdam Data Collective identifies five key steps within validation projects:

1) Model input
2) Model soundness
3) Model output
4) Regulatory compliance
5) Model implementation

We will use this structure to discuss the challenges that AI models could pose for validation teams and propose recommendations on how these could be solved.

Model Input

Model input testing generally consists of checking whether the data used for model development is complete, accurate and representative. The goal of this step in the validation process is to ensure that the model is trained and tested on valid data and will not perform poorly in application. One of the key capabilities that make AI methods so valuable is pattern recognition in large sets of (unstructured) data. This typically means that input datasets are larger and more complex than in traditional models. This poses a challenge for data quality checks, as it is harder to identify, classify and treat missing data and outliers. In a well maintained and structured Datawarehouse, information on data quality is readily available or relatively easy to deduct.

A data science modelling team, focused on increasing the predictive quality of their models, could use unstructured data, such as textual or visual data. This makes it significantly harder for validation teams to get a grip on the completeness, accurateness, and representativeness of the data. The validation team needs readily available expertise in fields such as Natural Language Processing, web scraping or computer vision to properly assess the model input. In this assessment it might be necessary to employ these techniques themselves.

A way for a validation team to properly validate data input in AI models is by focusing on the data quality strategy, rather than the data quality itself. Take the example of Money-Laundering detection in which the modelling team has set up a data-pipeline. The pipeline extracts data from sources within their Datawarehouse to publicly available info on social media platforms, transforms it to a suitable format and feeds it to the model. The model documentation should include a description of the various data sources, including types of data contained, the way it is extracted and transformed, known data issues, etc. Whether this is done properly is an important validation check in its own right. If done properly, the validation team only needs to compile a checklist and run some independent tests on samples of the input data.

Model Soundness

Model soundness aims to validate whether the model structure and assumptions are appropriate. AI models have the tendency to be convoluted, making it harder for an independent validator to grasp all aspects of the model. Apart from the specific knowledge and capabilities that this requires form the validation team, there are some practices that should be followed which makes the process of assessing the model soundness more manageable. First and foremost, it is important that the modeler makes use of widely used and accepted libraries and packages, e.g., *scikit-learn* and *Keras* when using python. By using components that are industry standard, the soundness of the separate model components can be assessed without going through complex, custom functions.

In addition, it is even more important to thoroughly document modelling design and assumptions, to ensure that important modelling decisions do not get buried in the complexity of the model itself. Naturally, it is key that the code is up to the highest standards. Clear and well-constructed code not only makes the code more readable but will also prevent coding errors. We recommend adding a proper git flow, naming conventions and code structure to the documentation requirements in the model governance framework.

Assumptions in an AI model can sometimes be hard to pin down. Using traditional models, such as linear regression, it is often clear what statistical assumptions are used. For instance, the errors are assumed to be randomly distributed over all samples. With AI models, this is less clear as often these models are not based on a statistical derivation, but rather on mathematical optimization procedures. Validators should therefore take extra care regarding the testing of assumptions during the development phase. Here, *Explainable AI (XAI)* is of high importance. This rapidly advancing field provides techniques which make it possible to relate input and output to each other independent of which algorithm has been used. Although some advanced AI techniques have a black-box character, techniques like *Shap* (Lundberg, 2021) and *Lime* (Reibeiro, 2021) can give a lot of insight in the predictions of the model, independent of its complexity. An effort to make the model explainable with XAI techniques should be made by the modelling team. It can be included in the model governance framework as requirement. In addition, the validation team can use these techniques to assess the soundness of the model.

In the example of an AML/CTF detection system, XAI can be used to interpret the model's predictions. It is essential that stakeholders like analysts and management are able to understand why certain transactions are marked as high risk. Not only does this cater to the needs of the several stakeholders, it also enables the validation team to open the black box of AI and assess model soundness.

Model Output

Model output testing validates whether the predictions produced by the model perform adequately. The validator checks whether the discriminatory power of the model and sub models are up to industry standard. The model output is also validated by checking the stability and robustness of the model estimates.

In general, with the range of machine learning algorithms that might be used in an AI solution, a wide range of metrics could be used to assess the discriminatory power of the model.

The more complex the data and the model becomes, the more important proper stability testing will be. We recommend implementing continuous stability testing to ensure the enduring quality of the model output. For example, the modeler could have implemented triggers that ensure early detection of potential incorrect results. If this aspect is lacking in the model documentation, it can be red-flagged.

Overfitting is a known pitfall for AI techniques. As both the amount of input data and the complexity of techniques used increases, the risk of overfitting becomes more serious. Techniques like K-fold cross validation should be employed to reduce the risk of overfitting. It is important to incorporate tests for overfitting in the validation framework.

The calibration accuracy and discriminatory power of the models should exceed the standards set for simpler models, otherwise it makes no sense at all to employ AI. This advantage in predictive power should be sustained over time and should be taken into account in the monitoring of the model. In model validation, the calibration accuracy and discriminatory power of models can be hard to test in hindsight. As mentioned before, it is essential that a proper split of train, test and validation data is implemented. Otherwise, validators are unable to independently determine the quality of the model, with a severe risk of overfitting.

In addition, the type of output of AI models may differ from those of more traditional models. In the field of AML/CTF detection , anomaly detection models can offer a great alternative or addition to business rule engines. Outlier detection is typically an unsupervised machine learning technique and can easily get complex. For example, we have seen banks employ auto-encoder networks, a deep learning model that aims to return the same output as input to flag outliers. Validating the output of such a model can be a challenge and might require expertise that is not readily available in typical model validation teams. Not only does the validator need expertise on regulatory compliance, data and modelling, they need to have specific knowledge on AI algorithms and tooling.

Regulatory Compliance

In the established field of financial risk modelling , regulatory compliance is a standard part of the model validation process. The European Banking Authority (EBA) publishes guidelines and regulatory standards to which financial institutions must adhere, although the exact interpretation can fuel discussions. A bank usually translates these guidelines into an internal regulatory checklist, which a validator can simply run down. However, in the case of an AI model, setting up a checklist for validation purposes can be tricky.

Creating a checklist for AI models is less straightforward than for the more well-trodden paths as the available guidelines are less well defined. However, there is some guidance for specific use cases. In the example of AML detection, both DNB and EBA guidelines offer some guidance on the use of AI. The relevant EBA guidelines are mainly aimed at the full framework of identifying and handling possible money laundering-related activities. There is a relatively high flexibility in the way that risks can be identified. DNB encourages the use of predictive modelling, but the guidelines leave room for interpretation. The EBA guidelines must be translated into a compliance checklist that provides a clear structure and framework to be followed.

In addition to readily available guidelines, the financial institution should develop its own views and policies on the use of AI. We recommend combining the available guidelines and internal policy and translating these to a use-case specific compliance checklist, such that there is a clear structure and framework that can be followed.

As there is less formal guidance for AI use cases, and thus more freedom in modelling choices, our approach is to make sure to involve the validation team early in the process, as visualized in figure 3. The development phase can be divided in a number of steps, which can be validated separately. By involving the validation team early and iterate often, you prevent valuable resources from entering dead-end streets. However, this should be done without compromising the independence of the validation results.

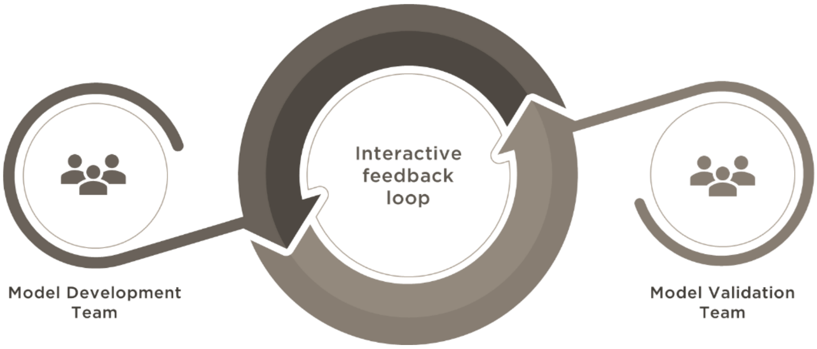

Figure 3: The more complex the model, the more important it is for the validation and development team to cooperate and interact. However, the validation needs to operate and form a judgement independently. A balance needs to be struck between cooperating effectively and independence of the validation.

Model Implementation

In our experience, implementation of AI models is a bottleneck in the model management cycle for many banks and companies. Proper implementation of an innovative model, including components like its data pipeline, continuous monitoring, and automatic calibration, requires a modern and flexible IT infrastructure. Implementing an innovative model and its components in a traditional infrastructure is often a road to nowhere. Most banks have realized that a drastic modernization of the infrastructure is needed, and are either modernizing their infrastructure, or setting up a completely new infrastructure in parallel. With these new infrastructures come new tools and concepts, which enable modelers and engineers to implement their models and plug them in the intended business process.

For validation, the flexibility and wide range of IT components used to implement the model raises a challenge. Proper validation of implementation requires knowledge of the IT components used. Naturally, part of this challenge can be mitigated by setting the right requirements in the model governance framework. By requiring the modeler to think about and document the intended implementation before or during the model development phase not only enables the validator to assess the implementation, it also increases the probability of the model being implemented and used successfully. Requirements in the model governance framework should cover aspects like data pipelines, necessary tooling, computational intensity, how the model results can be plugged into and used in the business process. Naturally, any modelling decision should be clearly documented and motivated, the code must be well documented and clearly structured. In addition, it is advisable to use libraries and algorithms that are industry standard and widely used. This enables the validator to assess the code both swiftly and thoroughly.

Model monitoring and change management is also part of model implementation. The challenge for validation will be in assessing the appropriateness of the procedures and whether aspects like automatic calibration and monitoring quality of the model are covered in the implementation. Naturally, validation becomes easier if the right requirements have been set beforehand.

For the example of AML/CTF detection, it is possible that an auto encoder network is able to perfectly filter out anomalies, but the model ends up not being used because there is no way to plug the outcome of the model into the transaction monitoring system in place. By setting the right requirements beforehand, efficiency in both the development phase and during the validation is ensured.

Conclusion

Undoubtedly, the use of AI models in the financial services industry will increase significantly in the near future. For the validation process, this will lead to additional challenges due to the lower interpretability of the techniques. To make proper and durable use of AI and prevent development teams from entering dead-end streets, it is important to reassess the existing validation frameworks. This article discusses how Amsterdam Data Collective approaches AI model validation.

This article on validation of AI models is based on our extensive experience in both developing and validating models . Given that the trend is to rely more and more on AI techniques, it will be increasingly important to have a suitable validation process in place. The issues addressed in this Insight should be viewed as a starting point in anticipating an increasing reliance on AI.

Do you want to know more or discuss this subject with our experts? Get in touch with Amsterdam Data Collective.

Authors

JAAP VAN ELSÄCKER

GOVERT VAN KONINGSVELD

JELMER QUIST

LAURENS STRONKS

JOOST VEENKAMP

Bibliography

- DNB. (2017). *Post-event transaction monitoring process for banks.*
- EBA. (2020). *EBA report on advanced analytics.*
- Lundberg, S. (2021). *SHAP (SHapley Additive exPlanations)*. Retrieved from Github repository: https://github.com/slundberg/shap
- Reibeiro, M. T. (2021). *Lime, Explaining the predictions of any classifier*. Retrieved from Github Repository: https://github.com/marcotcr/lime

We are ready for the future. Are you?

MAKE THE BEST DECISION, EVERY TIME

For many companies, generating and managing data in the era of big data feels like drowning in a sea of abundance. It's a constant challenge to understand customers better and stay responsive to their ever-changing needs. Faster innovation is called for: thinking beyond traditional frameworks to develop new services and business models.

Most organisations find they can't achieve this on their own. It takes relentless focus, the right expertise and an educated workforce to become a data-driven organisation. Overcoming this challenge is what Amsterdam Data Collective specialises in. We bridge the gap between strategy and data science. But data only becomes valuable when clients dare to let it shape their business. And trust us to join them on that journey

CALL FRANS BOSHUIZEN

+316 204 998 54
fBoshuizen@amsterdamdatacollective.com