Increasing Trust to AI in Finance: AI Model Validation Framework

Seçil Arslan[1,*] (Associate Partner), Buğra Akyüz[2] (Manager)

[1]Prometeia, River Plaza, Kat 19, Büyükdere Caddesi Bahar Sokak No. 13, 34394, Istanbul/TURKEY
[2]Prometeia, River Plaza, Kat 19, Büyükdere Caddesi Bahar Sokak No. 13, 34394, Istanbul/TURKEY

Abstract

Validation of Artificial Intelligence (AI) Models in the finance sector has been one of the most crucial phases of the AI models' life cycle. Although the finance sector is highly regulated and already familiar with validating traditional statistical methods in credit risk, they need an extension and adaptation to their as-is validation standards and frameworks for advanced AI algorithms. The extension is not only limited to credit risk but can also apply to divergent business domains. This paper highlights the risks of using AI in finance applications and provides significant motivations for having an AI validation framework to control and eliminate those risks. Besides, we underline the details of our framework's pillars by mapping them to well-known validation contexts like conceptual soundness, model performance, and model usage.

Keywords

Artificial Intelligence, Model Validation, Finance, Interpretability, Bias, Robustness, Fairness

1. Introduction

With the growth of the size of data and easy, low-cost accessibility to powerful processing units, the applications of Artificial Intelligence (AI) have been increasing tremendously in the finance sector. Although the finance sector is one of the early adaptors of programming technologies, it is still a blue ocean to use Machine Learning (ML) or AI-based systems and trust them in mission-critical applications.

It is crucial to recall the definition and life cycle of AI made by well-known references to support the requirement of having a full-fledged AI Model Validation framework.

Artificial Intelligence (AI) systems, according to OECD, are "machine-based systems with varying levels of autonomy that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions". AI techniques increasingly use massive amounts of alternative data sources [1]. They use machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes [1].

In another definition by [2]: "Machine Learning is programming computers to optimize a performance criterion using example data or past experience". A model is defined up to some parameters, and "learning" is the execution of a computer program to optimize the model's parameters using historical data. The model may be predictive to make predictions in the future, descriptive to gain knowledge from data or both.

Advanced AI approaches differ from traditional (statistical) approaches like Linear Regression or Logistic Regression. These traditional models are designed to make inferences about the relations between variables, following models and variables defined by human experts. These models can make reliable predictions, yet it is easier to interpret and explain them. As for ML/AI models, they are designed to make the most accurate predictions possible, as well as other inferences working on a data set and similar new data. They might sacrifice interpretability to increase their predictive power. Machine Learning, Deep Learning, Natural Language Processing (NLP), and Computer Vision are the primary fields of application for AI approaches.

Having been used more frequently in the banking sector recently, AI-oriented approaches seem to influence business strategies, risks, infrastructures, and operations of banks. For example; Decision Trees, Random Forests, Gradient Boosting Algorithms, and Neural Networks have started to replace models such as Logistic Regression as far as Credit Risk is concerned. In the domain of Operational Risk, Natural Language Processing methods concerning the manual entry of printed data and/or the classification of these data play a substantial role in the automation of these processes. NLP also contributes to the development of Chatbots and conversational interfaces for direct communication with clients. As far as financial fraud is concerned, AI approaches have significantly increased predictive power in terms of the detec-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
secil.arslan@prometeia.com (S. Arslan);
bugra.akyuz@prometeia.com (B. Akyüz)

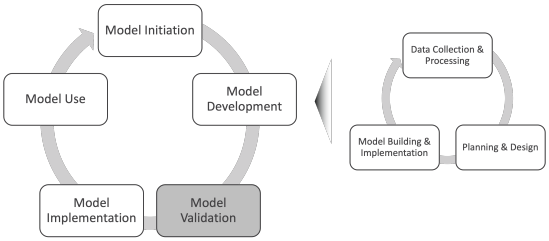

Figure 1: Model Life Cycle

tion of both credit card fraud and application fraud or Anti-Money Laundering owing to their capacity to model complex patterns within the data.

A typical life cycle of an ML/AI model has the following steps: (i) planning and design, data collection and processing, and model building and interpretation; (ii) verification and validation; (iii) deployment; and (iv) operation and monitoring [1] (Figure 1).

This paper aims to provide a full-fledged AI model validation framework that serves as a guideline to one of the critical life cycle phases of ML/AI models in the finance sector. Our approach discusses in detail the controls regarding **conceptual soundness** (model documentation, data validation, model design) and **model performance** suitability required for the entire life cycle process of an AI model as well as the controls necessary for the **model usage** in finance systems (production environment, usage and controls, monitoring approach). Those three concepts of validation framework refer to the 4 main pillars to be controlled and validated: Data, Methodology, Process, and Governance. Our framework aims to underline how to eliminate the risks of AI in finance applications by providing guides to validate models in terms of **data bias, quality, and privacy** issues; **robustness** and **fairness** of algorithms; preventing and detecting **overfitting** or **underfitting** performances and **interpretability** of ML/AI models and features.

2. Risks of AI in Finance

With the increasing number of ML/AI applications on credit risk, CRM/Marketing analytics, operational risk, process automation, fraud detection, and robo-advisory; financial companies need to take care of the risks of adoption of those ML/AI models in their as-is processes and workflows.

Due to high competition among financial institutions, many banks and insurance companies investing in applications of ML/AI in their core processes. However, the nature of ML/AI models depends highly on the selection of data samples, and learning from past data experience is the main driver of those algorithms. Also, unlike classical programming approaches, there is not a 100% expected outcome precision in those approaches. Although Banks are very familiar with model outputs that reflect a predictive approach, traditional well-known methods like Linear Regression and Logistic Regression are far more different than advanced ML techniques applied nowadays. Today, most of the ML approaches are more black-box and they require careful examination to create trust in robustness, fairness, data privacy, and bias concerns. In addition, unlike the linear methods, new algorithms require new methods to provide feature interpretability and model explainability like SHAPley or LIME [1].

The major risks of ML/AI models are about the responsibility and accountability of the models. The discussion is on who is accountable for unfair, biased results of a model; the historical data including biased information or the model developer not taking the necessary precautions, or the validation team not detecting the possible bias and fairness weaknesses.

Another risk is related to the typical problem of ML/AI models where they seem to perform very well in laboratory environments and cannot reflect the same performance in production environments. This may result from various reasons; one is that the data distribution or quality patterns may differ in production compared to the training data or the model may have overfitted on the training phase and no one has detected it.

All these risk factors affect the trust of ML/AI within institutions and compliance with legislation standards. Divergent applications including back, middle, and/or front-office related to credit, asset management, or even algorithmic trading [1] are at the core of those risks, and validation of these models has become the key point in managing the risks.

3. Motivation

In order to alleviate the risks that have already been mentioned, companies need a standardized guideline for validating AI models. Our main motivations for creating a validation framework are to (i) create trust for AI, (ii) guarantee compliance with legislation frameworks, and (iii) improve internal procedures.

First of all, improving the adoption of AI by creating "Trust" is a significant dimension of the need for an *AI Model Validation framework*. The adoption of AI in banking and finance applications is, although limited, increasing. Creating more awareness within the companies is possible by standardization of model validation processes that can fasten the early adoption of AI. Tier-1 banks prefer in-house developments of AI models, whereas, other banks may be limited in in-house capacities and prefer

vertical start-ups. Both cases require developing ”Trust” in the adoption of AI Models. Reports underline that 44% of models are in the pre-deployment phase and only 56% of them are in the deployment phase [3]. One way to improve trust is to validate the models before going live in production environments.

Secondly, it has become a good opportunity to create awareness and the need for an internal AI procedure (either developed or COTs services) that ensures AI Models are validated and checked with respect to legislations like EU ”AI Act” [4], or ”EBA Discussion Papers on Machine Learning for IRB Models” [5]. This will strengthen the arguments to convince customers of the need and empower ownership mechanisms.

Finally, although validation or audit teams of the AI Models are the main stakeholders of the AI Model Validation framework, they are not the only ones. The framework serves CDOs and CTOs of Banks to overview their available model development procedures. Risk awareness will be handled with a standard validation approach so that deployment processes will be smoother and aligned.

4. Approach

4.1. Validation Landscape

Our focus on the AI model's validation is to question and validate models including all steps from data creation to deployment [6]. Each step of a typical ML/AI model is subject to validation including data curation, training, adaptation, and deployment. Those concepts are represented in our framework in 4 main pillars of validation: (i) Data, (ii) Methodology, (iii) Processes, and (iv) Governance. Those validation steps respectively, validate models both in quantitative and qualitative perspectives.

The main actors within a model life cycle are model developers, validation teams, and end-users of the models (Figure 2). Each one has different roles and responsibilities in the life cycle of a model. Our framework primarily serves validation teams to support an internal defense mechanism within the company to check and approve the go/no-go decision of ML/AI models before going into production. Besides, the validation team is the main responsible to monitor regularly and repeat periodically some of the validation steps.

4.2. Validating on Three Concepts Mapped into Four Main Pillars

Most banks are more familiar with validation approaches and guidelines since the models are highly regulated by regulation and supervision agencies. (ECB, EBA). These agencies provide the rules and steps to be followed for the validation of regulative credit risk models. Since our

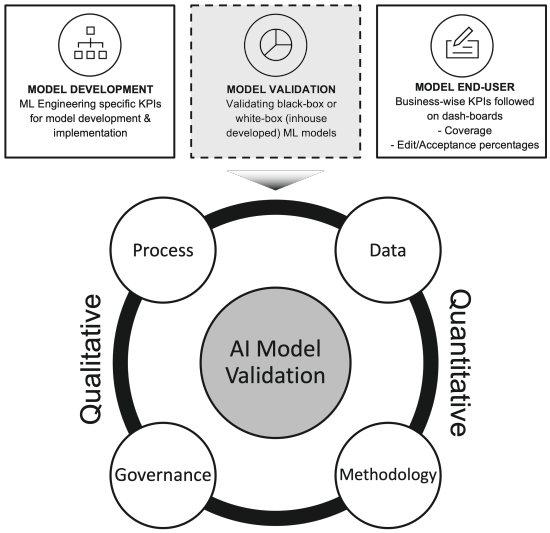

Figure 2: Validation Landscape

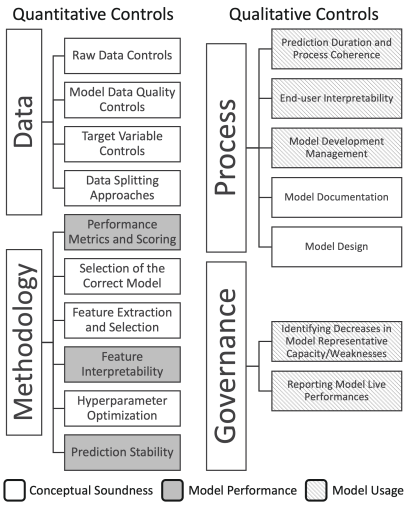

Figure 3: Four Pillars of Validation

AI Model Validation framework includes the validation of ML/AI models, including but not limited to, credit risk models, we mapped the new paradigms of validation to the as-is validation concepts that banks are already using internally [7].

Figure 3 demonstrates the mapping of our validation pillars into three concepts of validation.

Under the **conceptual soundness** dimension; the quality of model design, construction, and documentation is assessed. In addition to conventional steps like data validation, we need to focus specifically on ML-only

steps in methodology suitability: selection of the correct model, feature extraction/selection, and hyperparameter optimization. Even in data validation, we enlarge the typical validation phases of raw data, model data, and target variable quality controls with privacy and bias considerations on the selected data. Besides, the approach of choosing the correct data splits for utilizing the steps of training, parameter optimization, and testing the final results are questioned.

In the **model performance** part of the validation framework, model outputs are compared against the outcomes observed. In ML/AI, many different metrics and tests can be derived to quantify results. The important part is to provide guidelines to compare outcomes on the objective of models and define feature interpretability/-explainability with advanced methods. Our framework underlines the possibility of several different problem domains and algorithms that can be under validation. Each and every algorithm is questioned by choosing suitable performance metrics to compare results, feature selection and extraction methods, feature interpretability/explainability, and bias/variance concepts to detect models that are underfitting or overfitting.

The **model usage** phase focuses on not only validating outcomes and soundness but also the adaptation to processes and applications, the design of reflecting and digitizing processes with ML models, and also considers the "human-in-the-loop" strategy for the end-users. In addition, the model governance precautions, monitoring and reporting mechanisms, and continuous-learning techniques are examined.

4.3. Validation Types and Triggers

Validation procedures divide into two, according to their content and scope: Initial and Periodic. Initial validation is the end-to-end examination of a model after it has been developed and before it is released into the production environment. During initial validation, all controls under the four pillars mentioned above are completed. In periodic validation, however, changes in the population subject to the model and their effects on model performance are monitored in order to monitor the health of the model in general. Thus, it is aimed to detect models that are aging or whose performance is seriously deteriorated.

In both initial and periodic validation, the results of the tests applied are expressed by traffic lights. The green light indicates that the test has been passed, the red light indicates that the test has failed, while the yellow light indicates that the result is good enough, but can be improved. Since the question sets used for the validation processes consist of many questions under many categories, the use of traffic lights is important for these results to be clearly understood by relevant parties and for the final result of the validation to be determined depending on the rule sets defined on those traffic lights.

5. Conclusion

In this paper, we briefly discuss the definition of AI systems and their burgeoning usage in finance applications. We emphasize the possible risks of using AI in finance and underline the importance of the *validation phase* within the overall life cycle of a model. The driving factors behind preparing an end-to-end validation framework for AI models are the need for appropriate control over them and for creating trust in terms of bias, robustness, and fairness of the models.

Furthermore, we described different validation types and the logic of the traffic light approach for scoring models both initially and periodically in pre-deployment, and production environments, respectively.

Finally, our future works will focus on converting our AI model validation framework into an automated Software as a Service (SaaS) approach that is embedded into our Model Risk Management (MRM™) tool.

References

[1] OECD, Artificial intelligence, machine learning and big data in finance: Opportunities, challenges, and implications for policy makers (2021).

[2] E. Alpaydın, Introduction to Machine Learning, The MIT Press, 2020.

[3] B. of England, Machine learning in uk financial services (2022).

[4] E. Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021).

[5] EBA, Eba discussion paper on machine learning for irb models (2021).

[6] R. Bommasani, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021). URL: https://arxiv.org/abs/2108.07258. arXiv:2108.07258.

[7] E. C. Bank, Targeted review of internal models (2021).