Machine Learning Model: Perspectives for quality, observability, risk and continuous monitoring

Diego Nogare[1,2]; Ismar Frango Silveira[1]; Pedro Pinheiro Cabral[3]; Rafael Jorge Hauy[3]; Veronica Neves[3]

1 PPGEEC - Programa de Pós-Graduação em Engenharia Elétrica e Computação Mackenzie. São Paulo/SP, Brazil
2 ICTi - Instituto de Ciência e Tecnologia Itaú. São Paulo/SP, Brazil
3 Gestão de Validação e Riscos Operacionais - Itaú Unibanco. São Paulo/SP, Brazil

Email:{diego.nogare;ismar.frango;pedrop.pesquisa;rafael.hauy;veronica.neves88}@gmail.com

***Abstract*—The transition of machine learning (ML) and artificial intelligence (AI) projects from experimental stages to fully operational solutions presents substantial challenges. This is especially true for applications where these technologies play a critical role, demanding high-quality, reliable, and observable ML models. This paper explores the crucial aspects of continuous monitoring in ML models and emphasizes the need for a comprehensive approach that goes beyond technical development. It highlights that ensuring the reliability and robustness of deployed ML models requires a multifaceted framework encompassing data governance, model lifecycle management, and thorough team training. The paper addresses key aspects such as model quality, risk management, and the crucial role of observability in maintaining model stability and reliability in production environments. Using Itaú Unibanco as a case study, the paper showcases a robust model risk management approach and a dual monitoring system: an independent validation team oversees riskier models, while smaller models are monitored by their development team. The paper concludes by emphasizing the significance of a robust Model Risk Management (MRM) framework in the evolving landscape of AI and ML, particularly as these technologies become deeply integrated into various business operations. Highlighting that Itaú Unibanco's rigorous approach to model quality, observability, low risk, and continuous integration aligns with the regulatory requirements set by the Brazilian central bank.**

***Keywords*—Machine Learning, Model Monitoring, Model Observability, Model Quality, Model Risk Management**

I. Introduction

In the field of machine learning (ML) and artificial intelligence (AI), the promotion from experimental projects to production environment solutions is full of challenges. As these technologies become increasingly integral to critical applications across many industries, the imperative for maintaining low-risk or non-risk, high-quality, observable, and continuously monitored ML models has never been more explorated.

The adoption of ML and AI solutions for complex computational problems necessitates a paradigm shift in development processes. It is no longer sufficient to focus solely on the technical values of model development, there must be a concerted effort to engineer applications that operate seamlessly within dynamic production environments. This shift highlight the importance of robust model risk management, particularly in sectors like finance where the repercussions of model failures can be profound and has internal policies that regulate this practices. The potential for adverse consequences due to flawed models necessitates a comprehensive framework that encompasses not only technical development but also effective governance throughout the model's life-cycle. [1]

Such a framework mandates clear monitoring and control mechanisms to ensure ongoing model quality and observability. It aligns with regulatory expectations that model risk be treated with the same rigor as other operational risks, ensuring that institutions are proactive in identifying and mitigating potential issues [2]. The complexity of AI and ML models introduces new dimensions to risk management, further highlighting the need for robust practices. [1]

In this context, data accuracy and reliability become essentials, especially within data engineering functions integral to AI systems. As organizations increasingly rely on AI for automation and decision-making, the accuracy of data streams is critical [3]. While data ingestion processes aim for error-free ingestion, validating the content within those streams is equally crucial. Fast detection and fix of data issues are essential in AI-driven environments [3].

Therefore, as we stand at the edge of widespread ML and AI integration into critical applications, it is clear that ensuring low-risk, quality, observability, and continuous monitoring is not just a technical necessity but a comprehensive approach that covers data governance, model life-cycle management, and team training to reduce performance risks associated with AI

deployment [4], [5].

Itaú Unibanco bank upholds a robust model risk management process, ensuring adherence to clear definitions outlined in internal policies for all model development. This structured process includes audit reports assessing quality, risk, observability, and continuous monitoring using a risk-based approach.

II. Model Quality

The quality of machine learning models is multifaceted, in-depth several criteria beyond simple accuracy. A high-quality ML model has not only accuracy but also stability, resilience, a low computational and operational cost, and the ability to be trained or retrained quickly and reliably. Achieving these qualities in a ML pipeline often involves formulating the problem as a constrained multi-objective optimization task. There are some key aspects of model quality include, but not limited to: Small prediction error such as Root Mean Squared Error (RMSE); robustness to overfitting; fast computation time; stability and resilience. [6]

These quality criteria are essential for industrial applications of ML, which often rely on Machine Learning Operations (MLOps) pipelines to guarantee the long-term performance of models. The process of ensuring this versatile quality necessitates addressing data dependencies and mitigating ML-specific technical debt. [5], [6]

Robustness, a critical aspect of quality, specifically refers to a model's ability to remain unaffected by attackers or avoid misclassifications caused by outliers. Resilience is often used interchangeably with robustness. However, it's important to note that the definitions of reliability and resilience are less formally defined in ML compared to traditional software engineering. [7]

Data quality plays a crucial role in developing high-quality ML models. Ensuring the quality of input data streams by checking for missing values or anomalies, and employing appropriate preprocessing techniques like normalization, filtering, and so on, been essential steps in achieving reliable predictions. This focus on data quality is especially important in applications like image processing for manufacturing quality control, where training datasets might be small and differ significantly from the production images the model encounters during its operational life. [8], [9]

A. Model Risk Management

Model risk is the potential for possible adverse consequences appearing from decisions based on incorrect or misused model outputs and reports. The main regulatory document on model risk emphasizes that model risk should be treated with the same rigor as other risks. It also highlights the need to identify the sources of model risk and assess its magnitude to manage it effectively. [2]

Model Risk Management (MRM) covers the development of robust and consistent models, reliable implementation, appropriate use, consistent validation at an enough level of detail, and dedicated governance. It can also be understood as the process of mitigating risks associated with inadequate development, insufficient controls, and incorrect model use. MRM in the financial industry often involves expertise in technology, econometrics, and financial businesses. [2], [10], [11]

There are important consequences to effective MRM implementation in financial organization, such but not limited to:

Increasing reliance on models: Organizations rely heavily on models for decision-making, making it essential to manage the risks associated with them. [2]

Regulatory requirements: Regulators, particularly in finance, mandate that institutions implement MRM frameworks for all models used. For instance, the Federal Reserve's Supervisory Guidance on Model Risk Management (SR 11-7) provides guidelines for managing model risk. [1], [2] As of the Central Bank of Brazil, there is neither a regulation nor a guidance that establishes model risk management, some smaller Brazilian banks don't even have a team for this, although, there is the resolution nº 4557 that mandates a structure of risk management for banks, such as Operational, Market, Credit, Liquidity, and some other risks.

Financial losses: Errors in models can lead to significant financial losses, making MRM essential for protecting the organization's financial health. [2]

Reputational damage: Model failures can severely damage an organization's reputation, particularly in industries like finance where trust is paramount. [2]

Compliance and legal issues: Inadequate MRM can result in non-compliance with regulations and potential legal repercussions. [12]

As well, there are challenges in implementation of a successfully MRM framework, as but not restricted to:

Lack of standardized definitions: The definition of a model can vary among organizations, making it difficult to establish consistent MRM practices. [2]

Balancing accuracy and explainability: While advanced ML models can improve prediction accuracy, they often lack transparency, making it challenging to explain their decisions, which is unquestionable for regulatory compliance and stakeholder trust. [12]

Data quality and bias: Model accuracy heavily relies on data quality. Biased data can lead to discriminatory outcomes, raising ethical and legal concerns. [1], [12]

Model validation and monitoring: Continuously validating and monitoring models throughout their life-cycle is crucial to ensure their ongoing effectiveness and manage emerging risks. [1]

Shortage of skilled resources: Effective MRM requires specialized skills in areas like machine learning, statistics, and software development, which are often in high demand. [1]

Despite the challenges, organizations are increasingly adopting MRM practices, particularly with the rise of complex AI and machine learning models. Key initiatives include, but are not limited of:

Model inventory: A centralized repository containing information on all models used within the organization is must-have for effective MRM implementation. [2]

Model life-cycle management: A structured approach to managing models throughout their life-cycle, from development and validation to implementation and retirement, is essential. [2]

Independent model validation: Independent validation of models by teams separate from model developers helps ensure objectivity and identify potential flaws, and in some cases, its a legal demand from central bank. [2]

Model risk quantification: Developing methodologies to quantify model risk is nice to have for allocating capital reserves and making informed decisions about model deployment. [1], [2]

B. Model Monitoring

One key aspect of developing models is the monitoring of the model in production. One might be led to think, that if a proper train/test split is assured and other forms of data leakage were avoided, the quality of the model should already be guaranteed. However, other factors ranging from concept drifts to pipeline bugs and data quality issues can suddenly (or gradually) affect predictions.

Continuously monitoring machine learning models deployed in production environments is essential to ensure their sustained performance, reliability, and trustworthiness. This practice is crucial due to the dynamic nature of real-world data and the potential for model degradation over time, often caused by:

Concept drift: Changes in data patterns or relationships between variables over time; [13], [14]

Data drift: Shifts in the distribution of input data, leading to a mismatch between training and production data; [13]

Emerging scenarios: Unforeseen situations or inputs not encountered during model training. [14]

One of the main causes of model deterioration is data drift above described [15], but many other causes can lead to the phenomenon [16], [17], feature mismatch or data quality issues.

With all that in mind, it is important to define how the monitoring of the model should be structured. The first goal is to define a metric of quality for the model [18]. For labelled data this endeavour is considerably easier, the obvious choice is to use known metrics such as AUC-ROC, f1-score, accuracy, precision, recall and so on. The choice of the metric should take into account the kind of data being used, and the goal of the model prediction. Examples of reasonable metric choices would be:

Target imbalance: Databases with great target imbalance, such as fraud models, should avoid metrics that allow exploiting of the majority class, and instead focus on metrics that highlight performance on the minority class, such as precision, recall and F1-Score. [18]

Dramatic consequences for one output: Models with dramatic consequences for one output should ponder heavily whether true positives or true negatives are preferable. For email spam classification for example, the consequences of a false negative for a spam is the annoyance of having to manually sort it out. On the other hand, the consequences of a false positive for a spam is, potentially, missing and important email. Passing 10% of spam through might be more desirable than not passing 10% of through emails through. This should be taken into account when choosing metrics for evaluating performance. [18]

Unlabelled data: For unlabelled data, the metrics are harder to define and even more so for unstructured data, such as text or images. Human labeling of samples is a possible solution, but it is resource intensive in terms of personnel. Other metrics can be used such as volume of clusters, average distance to the centroid, proximity of answer to a knowledge database, etc. [15] [18]

Continuous monitoring acts as an early warning system, enabling timely interventions to mitigate performance degradation and maintain the model's effectiveness. This proactive approach helps avoid negative downstream impacts on decision-making and system functionality. The benefits of continuous monitoring could be, but not limited to:

Maintain model performance: Detect performance degradation early on, allowing for timely model retraining or adjustments to training data to ensure accuracy and reliability; [13], [19]

Ensure system reliability: Identify and address potential issues before they escalate into system failures or disruptions; [7], [19]

Enable informed resource allocation: Monitor system resource utilization to make informed decisions about infrastruc-

ture provisioning, whether using cloud-based or on-premise solutions; [19]

Facilitate responsible AI: Track model bias and fairness metrics over time to ensure ethical and responsible AI deployments; [13]

Support user trust: By ensuring consistent model performance and addressing potential issues, continuous monitoring helps build and maintain user trust in ML-powered systems. [20]

The increasing importance of continuous monitoring is highlighted by the growing adoption of the MLOps paradigm, which emphasizes the need to treat ML models as operational systems. MLOps encourages organizations to adopt practices and tools that streamline the deployment, monitoring, and maintenance of ML models, promoting their successful integration into production environments. [13], [19], [21], [22]

Numerous tools and techniques facilitate continuous monitoring, ranging from open-source libraries to commercial MLOps platforms. These tools offer features such as: [23]

Performance monitoring: Tracking metrics like accuracy, precision, recall, and F1-Score; [13], [19]

Data and concept drift detection: Identifying shifts in data patterns and model input distributions; [9], [13]

Bias and fairness monitoring: Assessing the model's fairness and potential biases over time; [13]

Alerting and Reporting: Generating alerts and reports to notify stakeholders of potential issues. [13]

By leveraging these tools and implementing a robust monitoring strategy, organizations can ensure the long-term performance, reliability, and trustworthiness of their deployed machine learning models.

C. Model Observability

Observability in Machine Learning, applied to production model pipelines, aims to provide enhanced visibility into the behavior of the ML system by leveraging telemetry collected during execution. The concept of observability extends beyond monitoring predefined metrics, which typically reflect only the overall system behavior. The proposition of observability is to enable practitioners to inquire about historical behaviors of the systems based on output data. [24]

The importance of observability in Machine Learning is related to the challenge of maintaining stability and reliability of models in production. Unlike during development, where there is real-time feedback and errors are more easily identifiable, models in production suffer from issues such as the lack of labels for predictions and silent failures that can occur at any stage of the model execution pipeline. [24]

The observability metrics, collected and analyzed by observability systems, provide valuable insights into the performance and behavior of Machine Learning models, enabling, but not limited to:

Bug detection: It is possible to identify anomalies in the model performance through the analysis of metrics such as accuracy, precision, recall, F1-score, AUC, and many others. Abrupt changes or deviations in these values may indicate issues in the pipeline, such as drift in the distribution of input data, data collection failures, or even drift of the model itself. [3], [24], [25]

Problem diagnosis: Through the analysis of observability metrics, it is possible to identify which components of the pipeline are contributing to the detected failures. [24]

Response and correction: Based on the analysis of observability metrics, it is possible to take actions to correct the identified problems, such as retraining the model with new data, adjusting hyperparameters, fixing errors in the code, or even revising the pipeline structure. [24]

There are three levels of observability of actions, ranging from highest to lowest: [26]

Fully Observable Action Sequence (FO): All actions in the plan appear in the observed action sequence. The observed action sequence contains all actions necessary to transition from each state to its corresponding successor state. This is the type of input trace accepted by all existing learning approaches. [26]

Partially Observable Action Sequence (PO): Some of the actions in the plan appear in the observed action sequence. At least one of the necessary actions from the plan is missing in the observed action sequence. [26]

Unobservable Action Sequence (NO): None of the actions in the plan appear in the observed action sequence. The observed action sequence is empty. [26]

III. Model Management at Itaú Unibanco

The methodology underpinning the deployment and evaluation of machine learning models at Itaú Unibanco is rooted in a rigorous, data-driven approach. Central to this methodology is the integration of machine learning models with the bank's existing IT infrastructure, which enables model deployment and continuous monitoring. Key components of the methodology include:

Development and validation: Utilizing a blend of historical data and real-time transactional data to train models, ensuring they are robust and capable of handling the dynamic nature of financial markets.

Model deployment: Implementing a phased rollout of machine learning models to monitor performance and mitigate potential risks.

Continuous monitoring and updating: Employing automated/semi-automated systems to track model performance against predefined metrics, indicating opportunities to retrain models using new data or adjust model parameters.

This methodological framework ensures that machine learning models deployed by Itaú Unibanco are not only tailored to the specific needs of the bank but are also flexible enough to adapt to changing market conditions.

Itaú Unibanco is committed to integrating advanced technologies, particularly machine learning, to enhance operational efficiency and customer service quality. The bank's operations span diverse financial services, including retail banking, wealth management, and corporate finance. Within this context, machine learning models serve as critical tools in several core functions:

Fraud Detection: Utilizing machine learning to identify and prevent fraudulent transactions in real-time, thereby reducing financial losses and strengthening customer trust.

Credit Scoring: Applying sophisticated algorithms to assess credit risk with higher accuracy, facilitating improved credit decision-making processes and lowering default rates.

Customer Service: Deploying chatbots and virtual assistants powered by machine learning to provide timely, relevant assistance, ultimately improving customer satisfaction.

Furthermore, machine learning plays a crucial role across various sectors. From retail banking for individual customers and small businesses, to wholesale banking for larger corporations, every division leverages machine learning to optimize processes, personalize services, and assess risks. Additionally, in the realm of investments, machine learning algorithms are employed to make informed decisions and predict market trends. In the finance department, this technology aids in managing assets and analyzing financial data efficiently. Even the legal sector benefits from machine learning by automating routine tasks and ensuring compliance with regulatory standards.

To ensure great model risk management, Itaú Unibanco has internal policies that regulate model development and usage, and these policies are constantly evolving to be as complex as models can be, for example: recently, with the revolution on Generative AI technologies, that is being quickly applied to a wide variety of use cases and requires a different approach to model risk management and monitoring, made the policies be adapted.

In a nutshell, at Itaú Unibanco, all models need to be added to the models' inventory before the use of it, so that the model governance team can assess the case, mitigate risks and give advice for the modelers. The models' inventory uses details from the model to classify each one inside a level of criticalness, for instance, if the model has direct impact on clients, regulations or the bank's balance sheet – high risk models need closer attention for the model risk management team than lower risk ones, such as an independent model validation before deployment and a more robust and more timely model monitoring.

The monitoring of models is a critical aspect of Itaú Unibanco operations. It involves selecting the appropriate metrics to track the performance of machine learning models accurately. Establishing a deterioration threshold for these metrics is essential to prompt actions when model performance declines. This proactive approach allows for timely interventions, such as retraining the model or implementing adjustments to maintain its effectiveness. Managing the implications of model performance degradation ensures that Itaú Unibanco upholds high standards of accuracy and reliability in its decision-making processes.

The definition of the monitoring metrics follows closely to the discussion of what metric is appropriate to evaluate a given model at all. The model development team sends, along with the model, a requisition for monitoring, where metrics are specified. The independent validation team can alter or question the specified metrics.

Along with the metrics a comparison model is also usually chosen. This comparison model is either an outside model that can be bought or a simpler model developed inside the Bank. After the metrics have been established and a comparison model is chosen, model risk management is carried automatically. If either of the following happens, a red flag is raised on that model for a given month:

- The model performance falls below a certain threshold, define by the independent validation team in conjunction with the model development team;
- The performance of the model is lower than the alternative model.

The adoption of machine learning models has yielded significant benefits for Itaú Unibanco, both in terms of operational efficiency and customer experience, and, for maintaining a pristine, robust, conservative image it has been building for the last more than one hundred years, this progress could only have been made with rigorous risk management processes.

IV. Final considerations

The field of Model Risk Management (MRM) is continuously evolving, influenced by technological advancements and regulatory changes. As AI and machine learning become more deeply integrated into business operations, the importance of robust MRM frameworks will continue to grow. The continuous monitoring of models can alert teams about data or concept drift, system reliability, low performance, and bias and fairness

situations. These alerts can be triggered before problems are noticed by customers. Analysis of observability metrics allows for a deeper understanding of how a machine learning system functions, the detection and diagnosis of problems, and more effective corrective measures. This contributes to a more robust and reliable lifecycle of machine learning model development, especially in critical systems. Itaú Unibanco addresses model management with a rigorous approach that covers model quality, observability, low risk, and continuous integration. This approach aligns with the legal requirements from the Brazilian central bank policies.

Acknowledgment

We would like to express our deep gratitude to Instituto Presbiteriano Mackenzie for the support provided during the development of this doctoral research. The financial support, infrastructure and resources made available were fundamental to the success of this work.

We also want to extend our thanks to the Instituto de Ciência e Tecnologia Itaú (ICTi) for the continuous encouragement and investment in Brazilian science. We firmly believe in the importance of your contribution to the advancement of knowledge and research in our country.

Any opinions, findings and explanations expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policies or position of Itaú Unibanco or Universidade Presbiteriana Mackenzie.

Referências

[1] S. Cosma, G. Rimo, and G. Torluccio, "Knowledge mapping of model risk in banking," *International Review of Financial Analysis*, p. 102800, 2023.

[2] D. S. Magalhães, S. B. S. Monteiro, and V. Vasconcellos, "Mitigation of model risk in a financial institution," in *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2022, pp. 1–7.

[3] J. F. Kurian and M. Allali, "Detecting drifts in data streams using kullback-leibler (kl) divergence measure for data engineering applications," *Journal of Data, Information and Management*, pp. 1–10, 2024.

[4] A. Bourgais and I. Ibnouhsein, "Ethics-by-design: the next frontier of industrialization," *AI and Ethics*, vol. 2, pp. 317–324, 5 2022. [Online]. Available: http://link.springer.com/article/10.1007/s43681-021-00057-0

[5] B. van Oort, L. Cruz, B. Loni, and A. van Deursen, "Project smells experiences in analysing the software quality of ml projects with mllint." Association for Computing Machinery (ACM), 5 2022, pp. 211–220. [Online]. Available: https://doi.org/10.1145/3510457.3513041

[6] E. Kannout, M. Grodzki, and M. Grzegorowski, "Considering various aspects of models' quality in the ml pipeline - application in the logistics sector." Institute of Electrical and Electronics Engineers Inc., 2022, pp. 403–412. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9908747&isnumber=9908601

[7] H. L. Franca, C. Teixeira, and N. Laranjeiro, "Techniques for evaluating the robustness of deep learning systems: A preliminary review." Institute of Electrical and Electronics Engineers Inc., 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9672592/

[8] P. Ruf, C. Reich, and D. Ould-Abdeslam, "Aspects of module placement in machine learning operations for cyber physical systems." Institute of Electrical and Electronics Engineers Inc., 2022. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9797080&isnumber=9797069

[9] B. Eck, D. Kabakci-Zorlu, Y. Chen, F. Savard, and X. Bao, "A monitoring framework for deployed machine learning models with supply chain examples." Institute of Electrical and Electronics Engineers Inc., 2022, pp. 2231–2238. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10020394&isnumber=10020156

[10] H. Jean-Baptiste, L. Tao, M. Qiu, and K. Gai, "Understanding model risk management–model rationalization in financial industry," in *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*. IEEE, 2015, pp. 301–306.

[11] H. Jean-Baptiste, M. Qiu, K. Gai, and L. Tao, "Model risk management systems-back-end, middleware, front-end and analytics," in *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*. IEEE, 2015, pp. 312–316.

[12] D. Chen and W. Ye, "Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring," in *Proceedings of the Third ACM International Conference on AI in Finance*, 2022, pp. 70–78.

[13] D. Nigenda, Z. Karnin, M. B. Zafar, R. Ramesha, A. Tan, M. Donini, and K. Kenthapadi, "Amazon sagemaker model monitor: A system for real-time insights into deployed machine learning models." Association for Computing Machinery, 8 2022, pp. 3671–3681. [Online]. Available: https://doi.org/10.1145/3534678.3539145

[14] I. L. Markov, H. Wang, N. S. Kasturi, S. Singh, M. R. Garrard, Y. Huang, S. W. C. Yuen, S. Tran, Z. Wang, I. Glotov, T. Gupta, P. Chen, B. Huang, X. Xie, M. Belkin, S. Uryasev, S. Howie, E. Bakshy, and N. Zhou, "Looper: An end-to-end ml platform for product decisions." Association for Computing Machinery, 8 2022, pp. 3513–3523. [Online]. Available: https://doi.org/10.1145/3534678.3539059

[15] C. Mougan and D. S. Nielsen, "Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap," in *AAAI Conference on Artificial Intelligence*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246294832

[16] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," *Knowledge-Based Systems*, vol. 245, p. 108632, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122002854

[17] J. Gama, I. Zliobaitè, A. B. abd Mykola Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, pp. 1 – 37, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:201087802

[18] T. Schröder and M. Schulz, "Monitoring machine learning models: a categorization of challenges and methods," *Data Science and Management*, vol. 5, no. 3, pp. 105–116, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666764922000303

[19] L. C. Silva, F. R. Zagatti, B. S. Sette, L. N. D. S. Silva, D. Lucredio, D. F. Silva, and H. D. M. Caseli, "Benchmarking machine learning solutions in production." Institute of Electrical and Electronics Engineers Inc., 12 2020, pp. 626–633. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9356298&isnumber=9356131

[20] H. Jayalath and L. Ramaswamy, "Enhancing performance of operationalized machine learning models by analyzing user feedback." Association for Computing Machinery, 3 2022, pp. 197–203. [Online]. Available: https://doi.org/10.1145/3531232.3531261

[21] R. Miñón, J. Díaz-De-Arcaya, A. I. Torre-Bastida, G. Zarate, and A. Moreno-Fernandez-De-Leceta, "Mlpacker: A unified software tool for packaging and deploying atomic and distributed analytic pipelines," 2022. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9854211&isnumber=9854207

[22] B. M. Matsui and D. H. Goya, "Mlops: A guide to its adoption in the context of responsible ai." Institute of Electrical and Electronics

Engineers Inc., 2022, pp. 45–49. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9808770&isnumber=9808570

[23] S. Idowu, D. Strüber, and T. Berger, "Asset management in machine learning: State-of-research and state-of-practice," *ACM Computing Surveys*, vol. 55, 12 2022. [Online]. Available: https://doi.org/10.1145/3543847

[24] S. Shankar and A. Parameswaran, "Towards observability for production machine learning pipelines," *arXiv preprint arXiv:2108.13557*, 2021.

[25] H.-L. Truong and T.-M. Nguyen, "Qoa4ml - a framework for supporting contracts in machine learning services," in *2021 IEEE International Conference on Web Services (ICWS)*, 2021, pp. 465–475.

[26] D. Aineto, S. J. Celorrio, and E. Onaindia, "Learning action models with minimal observability," *Artificial Intelligence*, vol. 275, pp. 104–137, 2019.