ISSN: 2473-4901
v11 n6338

AI & Machine Learning Deployment: Best Practices, Costs and Priorities

Nicholas Williams
williams.nicholas2100@gmail.com

Jeffrey Cummings
cummingsj@uncw.edu

Leo Gokce
gokcey@uncw.edu

Yu Wang
wangyu@uncw.edu

Cameron School of Business
University of North Carolina Wilmington
Wilmington, NC

Abstract

As artificial intelligence (AI) models become more prevalent across all fields, streamlined development of these models is becoming increasingly necessary. Deploying an AI model consists of integration with existing systems, monitoring various metrics related to the model, and maintenance of the model to keep it functional and up to date. Thus, successful deployment ensures value and sustainability. The objective of this research is to (1) identify the best practices within the phases of deployment, (2) explore cost requirements as well as strategies for savings, and (3) identify priorities during deployment. To explore these objectives, an exploratory study using semi-structured interview paradigm was developed and conducted with AI professionals with a qualitative analysis performed on the resulting transcripts. The analysis showed that participants emphasized explainable models that were accessible to users. Deployment costs were highly dependent on where the model was hosted and whether the model was developed in house or acquired from the commercial market. Finally, priorities were dependent on the type of model being developed, the users it would interact with, and the data it was handling. Regardless of these factors, all participants highlighted the importance of explainability, accessibility, and cost. These factors were prioritized by participants during model deployment.

Keywords: Artificial intelligence, machine learning, deployment priorities, best practices

ISSN: 2473-4901
v11 n6338

AI & Machine Learning Deployment: Best Practices, Costs, and Priorities

Nicholas Williams, Jeff Cummings, Leo Gokce and Yu WaNG

1. INTRODUCTION

Artificial Intelligence (AI) has quickly become technology's most enticing frontier. Driven by their ability to extract value from large volumes of data (Challoumis, 2024), AI tools have made this power more accessible across industries with increasing availability. However, despite its promises, AI implementation comes with its own set of challenges. One frequently cited challenge is rooted in the deployment of a trained model (Benbya, Davenport, & Pachidi, 2020; Paleyes, Urma, & Lawrence, 2023; Shankar, Garcia, Hellerstein, & Parameswaran, 2022). Even after successful deployments, it can be difficult to understand what a model needs to remain useful for an organization. This often requires monitoring and maintenance strategies that understand what to look for and how to respond effectively (Schober, 2022; Schröder & Schulz, 2022).

AI covers any time a machine tries to mimic human intelligence (Benbya, Davenport, & Pachidi, 2020), from rule-based expert systems to large language models. Technology in this category has been shown to make use of the copious amounts of data that are available in the modern world. AI demonstrates exceptional abilities in data analysis and augmenting human performance. It is able to assist with many repetitive workplace tasks and therefore, frees up people to take on more creative and innovative roles (Challoumis, 2024).

This paper focuses on AI, as well as a particular subset of AI classified as machine learning (ML). ML has numerous types of models that can be trained on data to make predictions. Ashmore et al. (2019) highlight the main steps in the ML process as follows: Data management, model learning, model verification, and model deployment. While data management, model training, and model verification are the foundation of a productive, well-performing model, this is not the end of the ML process. No matter what type of model is used, the deployment step is crucial to its initial and continued success.

While deployment is considered a crucial step, there have been a variety of approaches and suggestions to how this can be accomplished as well as the costs associated with it. This leads to the following research questions:

- How should companies effectively approach integration, monitoring, and maintenance of ML models?
- What costs are associated with implementing these practices and are they worth it?
- How should companies prioritize tasks within deployment?

2. BACKGROUND

ML, a powerful subset of AI that can be leveraged by many, is the practice of using algorithms to draw predictions from data. There are different types of ML (e.g., supervised, unsupervised, deep learning), yet they all follow the same basic steps as outlined by Ashmore et al. (2019). The first step is data management, which collects the data the ML model will be learning from. Next, the model learns to make predictions based on available data. Once satisfied with model predictions, the next step is model verification, which involves evaluating the model's performance on previously unseen data.

Once these steps are complete, the final step is deployment. Deployment involves 3 phases: (1) model integration, (2) performance monitoring over time, and (3) regular model maintenance and updates with new data. This study centers on deployment, with each phase discussed further in the subsequent sections.

Integration

Integration is the process of incorporating the ML model (i.e., the stored predictions) into the system it is to be providing information for. An organization must consider how an ML model will be integrated before attempting to deploy it. Even a well-trained and highly accurate model is useless if it cannot be incorporated into existing infrastructure. ML practitioners will encounter more infrastructure issues than expected during this initial deployment phase (Google Developers, n.d.).

A good first step is deciding whether the model will be making live predictions or producing outputs that can be stored and periodically updated. This can guide an organization to realize what type of infrastructure will best fit its needs. Zinkevich (n.d) also notes that once an infrastructure has been decided on, a model should be deployed to test how well it integrates. At the time of this initial deployment, value gain from predictions should not be prioritized (Google Developers, n.d.); instead, the focus should be on ensuring compatibility and seamless integration. This initial setup may require upskilling of existing employees or adoption of new technology that can make model predictions readily available to the application or data scientists (Benbya, Davenport, & Pachidi, 2020; Challoumis, 2024). Because of the cost both in time and money, organizations need to have an effective approach to this phase of deployment.

Monitoring

Once a model has been integrated, the next phase is monitoring. A common approach to monitoring the model is in-house. Schröder & Schulz (2022) present a variety of metrics that may be relevant to any given ML model: performance, robustness, confidence, economic, interpretability, and ethical metrics. Among these, performance metrics are most commonly used, providing insight into the accuracy of the model’s predictions. The type of model and what it predicts will determine the performance metric (e.g., accuracy, precision, etc.).

Monitoring model performance not only allows a user to see how well a model is doing, but can also be a strong counter to both problems related to drift (i.e., data and concept drift). Data drift is the concept that the data given to an ML model for it to make predictions will change over time (Ackerman et al., 2021). This is a very common occurrence within ML, as the world is ever changing. Concept drift is a phenomenon in which the performance of an ML model decreases over time, despite having the necessary up-to-date data (Schober, 2022).

While data drift is more specific to what data the model sees during training, concept drift describes a scenario in which the model was trained with a certain target in mind, but the target changes. This change in target could be due to a sudden change in the environment, a change in standards, or many other reasons. Concept drift is countered via retraining, just as data drift is. To deal with concept drift, the target must be defined in a way that demonstrates what is expected of the model. By seeing data that captures the changes to the desired outcome, the model can learn what predictions will be accurate for this new target (Schober, 2022). Thus, one area that needs to be explored is how practitioners approach drift across different domains.

Beyond the detection and prevention of drift, monitoring also has other utilities. Keeping track of how quickly a model’s performance degrades can enlighten ML practitioners to how often a model should be retrained, allowing for the creation of scheduled retraining. Keeping track of metrics that are important to an organization can also catch any mistakes a model may be producing before it goes live and has these errors reported by users. Even if a model appears to be performing correctly in its accuracy checks, an ethics check may reveal a bias that would damage the organization's reputation if customers were to see this.

Maintaining and Updating

Over time models can degrade (Patel, 2025), thus strategies to maintain or update a model must be considered. To update a model, two approaches are widely used: scheduled regular retraining and continual learning (Paleyes, Urma, & Lawrence, 2023). Scheduled retraining decides on a fixed amount of time between retraining the model with the latest data. This schedule should balance training frequency (enough to mitigate performance loss due to drift) and resource usage so that the retraining does not consume more resources than necessary.

Continual learning, by contrast, updates the model as it gets new data in. Many recommender systems use continual learning, as things change frequently with new data availability (Lee & Lee, 2020). The task an ML model is designed for will guide which type of updating it will work best with.

Cost

The value AI/ML provides can be very alluring, but it comes at a cost. There are many expenses that are outside the scope of this paper (e.g., data acquisition and storage, initial model training, personnel hours, etc.). The estimated cost of developing and implementing an AI/ML solution varies greatly from $10,000 to $1,000,000 (Shashkina, 2025). This large variation in price is largely chalked up to the complexity of the model and how long a company is willing to spend in the exploration phase.

Organizations have the choice to buy commercially available products or create their own models. There is also the decision of where to host, train, and update their ML model, either locally or via a cloud service. Each approach comes with benefits and issues, especially across the three phases of deployment.

Monitoring how much a model is costing is a wise and responsible step to take. If the ML model is being housed in a cloud vendor such as AWS or GCP, budgets can be set. These types of services usually provide a way to keep track of spending and offer automatic alerts. If the model is being housed on-site, metrics to keep track of compute time and other associated costs can be set up (Schröder & Schulz, 2022).

Deployment Best Practices and Priorities

Because of the vast differences discussed in the phases and costs associated with deployment, the goal of our study is to help develop the best practices and priorities that impact the cost and effectiveness of organizational AI/ML. Without proper implementation, anything the model provides becomes inaccessible and useless. Without proper monitoring, problems will go unnoticed, and the value provided by the model will begin to decline. By knowing what leads to smooth deployment operations, effective and enduring AI models can be incorporated anywhere.

3. METHODOLOGY

To address the research questions, a qualitative interview approach was used. Because of the challenges to capture the knowledge of experts in complex domains (i.e., AI) (Vasileiou, et al., 2018), the decision was made to focus on depth and conceptual understanding of the process by targeting specific roles (e.g., machine learning engineers) that may lead to more detailed and context-specific insights as opposed to interviewing everyone involved in AI (Turner & Hagstrom-Schmidt, 2022). Thus, the study focused on extensive interviews with 3 expert participants in the field with varying backgrounds and industries to provide diverse insights. Prior research has found that when participants are selected based on a high degree of expertise and role similarity, thematic saturation (i.e., identification of most major themes) may occur with as few as 3 interviews, especially when the topic is narrowly focused (Guest et al., 2006).

Participants

Participants were AI professionals with the responsibility of deploying AI/ML models. Participant 1 (P1) holds a managerial role with 3 years of AI experience, overseeing AI projects and a team of software engineers. Participant 2 (P2) is a data scientist with 2 years of professional AI experience. Participant 3 (P3) is an AI architect with 1 year of professional AI experience. All participants have directly contributed to the deployment of at least 2 AI projects.

Protocol

Interviews were conducted in a semi-structured fashion. Interview questions were constructed based on the research questions and best practices (i.e., open-ended, neutral, and clear (McNamara, n.d.; Turner & Hagstrom-Schmidt, 2022). The list of interview questions can be found in Appendix A. Interviews were conducted via Zoom, with an interaction time of 30-60 minutes. Conversations were not recorded for the privacy of the individual, but full transcriptions were collected and stored for analysis. Transcriptions were redacted of personally identifiable information and stored securely. This is based on similar interview paradigms implemented in past research (Shankar et al., 2022).

Analysis

The standard for qualitative research analysis is transcript coding (i.e., qualitative content analysis) (Shankar et al., 2022). This practice extracts common themes across interviews. MaxQDA, a commonly used qualitative data analysis software, was employed in this study. Deidentified transcripts recorded from Zoom were imported into the software. Coding passes were then performed on each transcript using a top-down approach. 8 codes were derived from the research questions and literature and applied to relevant segments within each interview. A list of codes can be found in Table 1. In total, 139 segments across the 3 interviews were given codes. Common themes, unique approaches, and surprising contrasts are presented in the following section.

Codes	Segments	%
Cost	39	28.06
Priorities	24	17.27
Maintenance	18	12.95
Monitoring	18	12.95
Integration	16	11.51
Demographic	10	7.19
AI process	8	5.76
Strategy & Governance	6	4.32

Table 1. Qualitative Content Analysis Codes

4. RESULTS & DISCUSSION

In this section, the results of the qualitative content analysis are presented and discussed. First, best practices related to the phases of deployment are covered. Next, cost considerations and management strategies are discussed. Finally, we elaborate on what should be prioritized when looking to deploy AI/ML models.

Integration

The goal of integration is to make the AI/ML model or its outputs accessible to users. In the case of an AI chatbot, this would be making sure users are able to speak with it, whereas for a sales forecasting ML model, this would be making sure the predictions can be seen by the stakeholders to make relevant business decisions.

Participants identified integration as a critical first step in AI/ML deployment. Participants specifically focused on the explainability of the models, which refers to the ability to understand how a model generates its outputs. It did not matter if an interviewee was referring to a pre-trained AI chat model or an in-house trained ML prediction model; understanding how an output was reached was always cited as an important factor.

"You don't really want to focus in on one variable when you're explaining the model, you kind of want to tell a story about all the significant variables at once." P2

P2 continued to describe how understanding a model's thought process can easily translate into providing logical, data-driven justifications for model outputs that domain experts can understand and agree with. In this case, ML models are strong tools for pattern recognition to assist human judgment. P2 also valued explainability above model performance, indicating that performance can be increased if the model can be explained, but a model that cannot be explained will be much more difficult to improve. This sentiment concerning explainability was echoed by the other participants. P1 cited tools like LangGraph and LangSmith for their ability to demystify a model's chain of reasoning. Being able to see where things went wrong allows the developers to adjust that parameter in a way that steers the model toward the desired outcome.

P2 was the only participant to talk about security. This is an important factor to consider, especially during integration. This is the phase where a model is about to be accessible to more than just the data scientists. Many AI/ML projects deal with potentially sensitive data. The handling of the data being fed to the model must be secure, and the outputs of the model must also be secure. Secure practices should be implemented every step of the way, including during deployment.

Other best practices mentioned by the participants included combining models to make an ensemble model to increase decision confidence and performing stress testing to evaluate system resilience under expected amount of traffic during integration.

Monitoring

As previously mentioned, performance is a metric that predictive models use to know how well they are doing their task. While this has been cited in previous research as the most important due to its direct link to the model's value, participants in our study claim that explainability metrics are even more important. Knowing how a model arrives at its final output allows data scientists to give raw data to support correct decisions, as well as debug incorrect decisions.

Other features that are directly tracked include: how long users interact with chatbot AI models, latency of model response, token usage, and cost. For indirect monitoring, P1 explains an interesting process in which an AI model is integrated; users provide feedback on their experience with it, then the development team recruits a separate AI model to perform sentiment analysis on user feedback. This allows a way to quickly get a feeling for what is and is not meeting users' expectations, allowing for rapid fixes that satisfy people who use the AI model. P3 mentions that they track what users are engaging with their AI chatbot model for. This is done to identify any common tasks that users may frequently ask the model to perform. Once those are identified, the model can be tuned to handle those common tasks more efficiently, decreasing compute costs.

Maintenance

Maintenance covers the steps needed to keep a model working properly, up-to-date, and expanding functionality. Just like all technology, new pre-trained, commercially available models are regularly being released. Participants stated that they did not always immediately update to the latest model. Instead, they evaluate if the increase in performance is worth the increase in price. If not, they continue to use the slightly older model that is still providing satisfactory performance. Conversely, it was also mentioned that money can potentially be saved by updating to the latest model if the performance and cost differences of a new model warrant the switch.

P1 also mentioned the usage of an "automated control test suite”. This is an automated test for a group of use cases for which the model should provide accurate results every time. It is used any time an update is made to ensure that basic functionality has not been broken.

For in-house models, participants mentioned that models should be retrained regularly to avoid any type of drift, with retraining dependent on the task being performed. P2 gave the example of a model that is used to assist in stock trading, which should be updated daily, at a minimum. This could be juxtaposed to a model that is used for annual sales forecasting, which may only need to be retrained once per quarter. P2 also discussed backtesting for model updates. This process trains a model on historical data. The model is then tested on more historical data so its performance can be immediately evaluated. This is useful in maintenance because it would be bad if an updated model were integrated but then performed worse than its previous iteration.

Backtesting is a way to vet an updated model before presenting it to end users. P3 indicated that the models they built did not yet have a need for maintenance. The models were all under 3 months old and used for tasks where new data was not greatly different than old data. This should factor into developing a maintenance strategy.

Deployment Best Practices Summary

Based upon the interviews conducted, the following summarizes the results across the phases of deployment. Organizations should approach integration, monitoring, and maintenance of models by focusing on accessibility, explainability, and security, while tailoring strategies to specific use cases.

Explainability should be prioritized during integration to help stakeholders understand and trust the decisions generated by the models. Security during integration safeguards sensitive data, while advanced practices like ensemble modeling and stress testing improve reliability. Monitoring tracks many different aspects, such as performance metrics, user feedback, and engagement patterns, to refine models and optimize their functionality.

Explainability, again, plays a key role in debugging and decision-making, surpassing the importance of performance alone. Maintenance strategies should involve regular retraining to prevent drift, automated testing to verify functionality, and cost-effective evaluation of updates to pre-trained models. By adopting these approaches, companies can ensure that their AI models remain valuable decision-making tools.

Cost

Cost is one of the main factors that a business will consider when using AI models. The interview participants had very diverse approaches to powering their companies with AI, leading to very different allocations of resources. The two factors that determined where money was focused the most were the origin of the model and how the model was hosted.

Models can originate from within the company or be acquired from an outside vendor (i.e., build vs. buy). P1 uses pre-trained, commercially available LLM models that can be tuned and adjusted to the specific task they require. The main justification for choosing to buy instead of build was that the AI landscape is changing at a pace that is very difficult to keep up with:

“We're subject to the leapfrog effect, right? So, by the time you have invested the time and resources to train a model, the commercially available models have already bypassed you … again and again, we've seen that happen and we've seen competitors try [to keep up with] that and then fall short.” P1

Using commercially available models was described as “pay as you go”. Cost scales with token usage, latency/speed of response, and amount of data transferred. One benefit of this approach is that most of the cost is upfront. Once the model has been paid for, monitoring and maintenance are inexpensive, as they are provided by the vendor.

Another benefit of buying a commercially available model is that it is ready for production much sooner. Many commercial models have out-of-the-box capabilities, providing value as soon as they are purchased. This trade-off of high up-front cost with immediate usage vs the lower costs but slower time to market of models built in-house is one to consider. Opting to buy commercially available models appears to be worth considering if the business plans for scaling up in size over time would benefit from having the most recent models available, or if immediate responses are necessary.

In contrast, P2 chose to build models in-house instead of buying commercially available models. Importantly, these in-house models were predictive ML models, not LLMs like those used by P1. P2 claims that integration is cheaper, while maintenance is more expensive. This is in direct contrast to P1, who said their greatest expense was in integration, and maintenance was not very costly. Since P2 does model retraining using cloud computing, the computing cost must be paid every time a model is updated.

P3 has an entirely different experience with cost, as they buy hardware, completely avoiding cloud computing costs. There is a cost in acquiring the hardware which must be covered every time the organization scales up by adding a new AI project. However, by training and maintaining their models locally, they do not have to pay every single time they want to update their models. This approach is best suited for those who can acquire hardware cheaply or who do not plan to have a vast amount of AI products in their organization.

Participants suggested an artful balance must be struck between processing power and time. This is especially true for those who decide to use cloud computing to power their AI models. Better processing can complete tasks in a shorter amount of time, at the cost of a higher rate. If time is not a critical factor, less powerful processing can be used, resulting in a cheaper rate, but increased time to complete the task.

Participants also mentioned multiple tradeoffs where there was potential to save money. P1 leverages batch jobs to save money when time is not a critical factor. Thus, if you can pay over time, you do not need to spend as much money. Another option to consider is looking for open-source software. As the AI landscape develops daily, more and more solutions are becoming available. P1 cited this as a strategy that is considered when possible and reported that money was saved when these solutions were employed.

	P1	P2	P3
Model origin	Commercial	In house	In house
Hosted	Cloud	Cloud	Local
Most expensive phase	Integration	Maintenance	Integration

Table 2. Participant Experience & Cost

Priorities

There are many variables to consider when looking to implement an AI project. The first to consider is whether the model will be built in-house or outsourced via a commercially available model. Important factors to consider when making this decision include current availability of resources, the pacing of the development team, and how important accessibility is.

As previously mentioned, using commercially available models incurs greater cost up front. If this can be afforded, paying for a commercially available model is a viable route. It has also been mentioned that the AI world is constantly evolving, making it difficult to keep up with. If the business must be using the most advanced, cutting-edge AI models, they will either need to have a team to support this rapid development cycle or turn to commercially available models. The availability of models is a unique problem that was only mentioned by P3. This interviewee works in a remote, rural area where natural disasters frequently cut off communication to the outside world. Since some of the models built here provide important information, cloud hosting was not an option. This is an important reminder that any models that have outputs or interactions with critical systems must be available and not reliant on the cloud.

There are some common goals for models, regardless of the domain or particular use case they are applied to. These include successful integration, explainability, alignment with business goals, and security. Integration has been covered extensively, but the importance of users being able to interact with the model or its outputs cannot be understated. Explainability has also been discussed in depth, as participants highlighted it as the most important factor of a deployed model. To have an explainable model means that undesired outputs can be traced to the point of failure and subsequently adjusted.

Furthermore, if a model’s steps can be traced, that can be translated into valuable information that non-technical members can benefit from hearing. From here, priorities become domain-specific. If a model is going to be interacting with customers, latency must be considered, as they do not tolerate slow response times, according to participants. Alternatively, if a model is only going to be used internally, response speed may not be as great a concern.

The domain in which the model will be operating also provides context for maintenance. This was a difference observed between two of the interviewees. P1 had a customer-facing model that needed to be kept up to date. In contrast, P3 had internal models that did not work with data that changed frequently. This led P1 to prioritize maintenance more than P3, who said, “Model drift is not a … concern for us currently.” They continued to explain that due to the invariability over time of their data, there is not much pressure to regularly retrain or update models.

Security is a concern that must be addressed at every phase of a model’s life. participants briefly mentioned security, mostly related to ensuring the data is only accessible to the necessary parties. Security must also be checked to ensure that the model is not able to communicate any sensitive information to people who should not have access to it.

Automation is an important part of AI models, as it saves a lot of time and keeps things up to date. When first releasing a model, automation does not have to be a priority, as the model will still serve its purpose without automatic updates. Participants described how this practice is not vital for release, but will quickly become important, as the model’s lifecycle continues.

With these priorities considered, a strategy can be devised that will promote initial success and provide steps for a model to have a long, sustainable life. Some priorities are dependent on what domain the model is operating in, such as latency, which is important for customer-facing models. However, successful integration and being able to explain how a model got to its output are steps that are crucial for success, regardless of the operating domain.

5. CONCLUSION

In this study, results were presented from a semi-structured interview of AI deployment practitioners. Findings suggest is paramount, while monitoring and maintenance are later, yet still important concerns. Costs within deployment were dependent on whether a model was built in-house or by a commercial vendor. In-house models result in a slow increase in price when hosted via a cloud provider, as maintenance and updating incur compute costs. Commercially available models have the bulk of the cost upfront, as maintenance and updating are handled by the provider. Priorities for any type of model revolved around accessibility, explainability, and cost.

Overall, the study provides valuable insights into the practical aspects of AI/ML deployment and identifies approaches for organizations looking to implement these technologies effectively. The findings contribute to the existing literature by providing a detailed analysis of real-world experiences and challenges faced by AI/ML professionals.

6. FUTURE WORK AND LIMITATIONS

Future work could build on this work by examining the power of explainability and researching strategies to ensure this is achieved. This study was limited to a small sample size. Future studies could expand on these ideas across multiple industries. A broader scope could be used to reinforce the findings presented.

A framework for necessary and highly important steps and decisions within deployment could be developed. Part of this framework could include AI auditing, which involves reviewing algorithms for fairness, compliance, accountability, etc. While this topic was beyond the scope of this study, it should be researched and considered in a responsible deployment framework.

Security could be researched as it relates to the steps of deployment, with secure practices laid out for practitioners. Security may also fit into the previously mentioned potential concerns about using commercially available models. If sensitive data must be communicated to third-party vendors, there could be potential for vulnerabilities to arise.

This study was limited to interviews of the experiences of a few people. Future research could take a quantitative approach, especially when examining cost. This could incorporate the data of a great number of individuals and report concrete numbers. By addressing these areas, future research can further enhance our understanding of AI/ML deployment and contribute to the development of more effective and sustainable AI solutions.

7. REFERENCES

Ackerman, S., Raz, O., Zalmanovici, M., & Zlotnick, A. (2021). Automatically detecting data drift in machine learning classifiers (arXiv:2111.05672). arXiv. https://doi.org/10.48550/arXiv.2111.05672

Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges (arXiv:1905.04223). arXiv. https://doi.org/10.48550/arXiv.1905.04223

Benbya, H., Davenport, T., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. MIS Quarterly Executive, 19, 9–21. https://doi.org/10.2139/ssrn.3741983

Challoumis, C. (2024, October). The economics of AI - How machine learning is driving value creation. https://doi.org/10.5281/zenodo.13929032

Deepchecks Community. (2025, February 6). Model versioning for ML models: A comprehensive guide. Deepchecks. https://www.deepchecks.com/model-versioning-for-ml-models/

Google Developers. (n.d.). *Rules of machine learning: Best practices for ML engineering*. Google. Retrieved June 15, 2025, from https://developers.google.com/machine-learning/guides/rules-of-ml

Guest, G., Bunce, A., & Johnson, L. (2006). *How many interviews are enough? An experiment with data saturation and variability*. Field Methods, 18(1), 59–82. https://doi.org/10.1177/1525822X05279903 Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. The Lancet Digital Health, 2(6), e279–e281. https://doi.org/10.1016/S2589-7500(20)30102-3

McNamara, C. (n.d.). General guidelines for conducting research interviews. Retrieved June 1, 2025, from https://management.org/businessresearch/interviews.htm

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1–29. https://doi.org/10.1145/3533378

Patel, H. (2025, February 11). ML model deployment challenges. Censius. https://censius.ai/blogs/challenges-in-deploying-machine-learning-models

Schober, A. (2022, September). *What is concept drift and how to detect it*. Motius. Retrieved [June 1, 2025], from https://www.motius.com/post/what-is-concept-drift-and-how-to-detect-it.

Schröder, T., & Schulz, M. (2022). Monitoring machine learning models: A categorization of challenges and methods. Data Science and Management, 5(3), 105–116. https://doi.org/10.1016/j.dsm.2022.07.004

Serban, A., van der Blom, K., Hoos, H., & Visser, J. (2020, October). Adoption and effects of software engineering best practices in machine learning. In Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (pp. 1–12). https://doi.org/10.1145/3382494.3410681

Shankar, S., Garcia, R., Hellerstein, J. M., & Parameswaran, A. G. (2022, September 16). Operationalizing machine learning: An interview study (arXiv:2209.09125). arXiv. https://doi.org/10.48550/arXiv.2209.09125

Shashkina, V. (2025, February 11). Machine learning (ML) costs: Price factors and real-world estimates. ITRex. https://itrexgroup.com/blog/machine-learning-costs-price-factors-and-estimates/

Turner, D. W., III, & Hagstrom-Schmidt, N. (2022, January). Appendix: Qualitative interview design. https://odp.library.tamu.edu/howdyorhello/back-matter/appendix-qualitative-interview-design/

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. BMC Medical Research Methodology, 18(1), 148. https://doi.org/10.1186/s12874-018-0594-7

Yasenchak, E. (2025, February 3). The challenges of deploying machine learning models: Best practices. Medium. https://medium.com/@emyasenc/the-challenges-of-deploying-machine-learning-models-best-practices-7c616a5a07d2

Zinkevich, M. (n.d.). *Rules of machine learning: Best practices for ML engineering*. Google Developers. https://developers.google.com/machine-learning/guides/rules-of-ml

APPENDIX A.

Semi-Structured Interview Questions

1. What type of AI do you use?
2. Can you give a general overview of your AI process from conception to deployment? (A, B)
3. How do you implement model predictions into your system? (A)
4. Do you track any aspects of your AI after it has been deployed? (A)
    a. What metrics do you use to monitor each aspect? (A)
5. Do you update your AI? If so, how? (A)
    a. What prompts the need for an update? (A, B)
6. Have you used any techniques that ended up being a waste of resources in hindsight? (A, B, C)
7. Were there any practices/tools/strategies that were costly to implement (C), but worth the spending in the long run? (A, B, C)
8. Could the model be live in production without one or more of these steps? (A, B, C)
    a. Would you see the same amount of value if those steps were skipped?
9. What are the priorities when deploying an AI model? (B)