# Challenges faced at each stage in Deploying Machine Learning: A Study

SADHANA PANDEY , POOJA KARMA
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SAGE UNIVERSITY, INDORE (M.P)

**ABSTRACT** *Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Our survey shows that practitioners face challenges at each stage of the deployment. The goal of this paper is to layout a research agenda to explore approaches addressing these challenges.*

**KEY WORDS***: Machine Learning (ML),artificial intelligence(AI),data oriented architecture(DOA),Natural Language Processing(NLP), Hyper parameter optimization(H.P.O).*

## 1 INTRODUCTION

According to a recent global survey conducted by McKinsey, machine learning is increasingly adopted in standard business processes with nearly 25 percent year-over-year growth [1] and with growing interest from the general public, business leaders [2] and governments [3]. This shift comes with challenges. Just as with any other field, there are significant differences between what works in academic setting and what is required by a real world system. Certain bottlenecks and invalidated assumptions should always be expected in the course of that process. As more solutions are developed and deployed, practitioners sometimes report their experience in various forms, including publications and blog posts.

In our survey we consider three main types of papers:

- Case study papers that report experience from a single ML deployment project. Such works usually go deep into discussing each challenge the authors faced and how it was overcome.
- Review papers that describe applications of ML in a particular field or industry. These reviews normally give a summary of challenges that are most commonly encountered during the deployment of the ML solutions in the reviewed field.
- “Lessons learned” papers where authors reflect on their past experiences of deploying ML in production.

This survey supports our goal to raise awareness in the academic community to the variety of problems that practitioners face when deploying machine learning, and start a discussion on what can be done to address these problems.

## 2. THE ML WORKFLOW

According to Ashmore et. al. [11], the process of developing an ML-based solution in an industrial setting consists of four stages:

- Data management, which focuses on preparing data that is needed to build a machine learning model;
- Model learning, where model selection and training happens;

- Model verification, the main goal of which is to ensure model adheres to certain functional and performance requirements; 2
- Model deployment, which is about integration of the trained model into the software infrastructure that is necessary to run it. This stage also covers questions around model maintenance and updates.

Each of these stages is broken down further into smaller steps. It is important to highlight that the apparent sequence of this description is not necessarily the norm in a real-life scenario. It is perfectly normal for these stages to run in parallel to a certain degree and inform each other via feedback loops

# 3. DATA MANAGEMENT

Data is an integral part of any machine learning solution. Overall effectiveness of the solution depends on the training and test data as much as on the algorithm. The process of creating quality datasets is usually the very first stage in any production ML pipeline.

## 3.1 Data collection

Data collection involves activities that aim to discover and understand what data is available, as well as how to organize convenient storage for it. The task of discovering what data exists and where it is can be a challenge by itself, especially in large production environments. Finding data sources and understanding their structure is a major task, which may prevent data scientists from even getting started on the actual application development.

## 3.2 Data preprocessing

The preprocessing step normally involves a range of data cleaning activities: imputation of missing values, reduction of data into an ordered and simplified form, and mapping from raw form into a more convenient format. Methods for carrying out data manipulations like this is an area of research that goes beyond the scope of this study.

## *3.3* Data augmentation

There are multiple reasons why data might need to be augmented, and in practice one of the most problematic ones is the absence of labels. Real-world data is often unlabeled, thus labeling turns out to be a challenge in its own right. We discuss three possible factors for lack of labeled data: limited access to experts, absence of high-variance data, and sheer volume. Labels assignment is difficult in environments that tend to generate large volumes of data, such as network traffic analysis.

## 3.4 Data analysis

Data needs to be analyzed in order to uncover potential biases or unexpected distributions in it. Availability of high quality tools is essential for conducting any kind of data analysis. One area that practitioners find particularly challenging in that regard is visualization for data profiling [13]. Data profiling refers to all activities associated with troubleshooting data quality, such as missing values, inconsistent data types and verification of assumptions. Despite obvious relevance to the fields of databases and statistics, there are still too few tools that enable efficient execution of these data mining tasks.

# 4. MODEL LEARNING

Model learning is the stage of the deployment workflow that enjoys the most attention within the academic community. All modern research in machine learning methods contributes towards better selection and variety of models and approaches that can be employed at this stage.

, there is still plenty of practical considerations that affect the model learning stage. In this section, we discuss issues concerning three steps within model learning: model selection, training and hyper-parameter selection.

### 4.1 Selection of model

In many practical cases the selection of a model is often decided by one key characteristic of a model: complexity. Despite areas such as deep learning and reinforcement learning gaining increasing levels of popularity with the research community, in practice simpler models are often chosen as we explain below. Such model include shallow network architectures, simple PCA-base approaches, decision trees and random forests.

### *4.2* Training

One of the biggest concern with model training is the economic cost associated with carrying the training stage due to the computational resources required. This is certainly true in the field of natural language processing (NLP), as illustrated by Sharir et al. [14]. The authors observe that while the cost of individual floating-point operations is decreasing, the overall cost of training NLP is only growing. They took one of the state-of-the-art models in the field, BERT [15], and found out that depending on the chosen model size full training procedure can cost anywhere between $50k and $1.6m, which is unaffordable for most research institutions and even companies.

### 4.3 Selection of Hyper-parameter

In addition to parameters that are learned during the training process, many ML models also define several hyper-parameters. Hyper-parameter optimization (HPO) is the process of choosing the optimal set of these hyper-parameters. Most HPO techniques involve multiple training cycles of the ML model. Besides, the size of HPO task grows exponentially with each new hyper-parameter, because it adds a new dimension to the search space. As discussed by Yang and Shami [40], these considerations make HPO techniques very expensive and resource-heavy in practice, especially for applications of deep learning. Even approaches like Hyperband [17] or Bayesian optimization [18], that are specifically designed to minimize the number of training cycles needed, are still not able to deal with certain problems due to the complexity of the models or the size of the datasets

## 5. VERIFICATION OF MODEL

The goal of the model verification stage is multifaceted, because an ML model should generalize well to unseen inputs, demonstrate reasonable handling of edge cases and overall robustness, as well as satisfy all functional requirements. In this section, we discuss issues concerning three steps within model verification: requirement encoding, formal verification and test-based verification.

### 5.1 Encoding of requirement

Defining requirements for a machine learning model is a crucial prerequisite of testing activities. It often turns out that an increase in model performance does not translate into a gain in business value, as Booking.com discovered after deploying 150 models into production [19]. Therefore more specific metrics need to be defined and measured, such as KPIs and other business driven measures. In the case of Booking.com such metrics included conversion, customer service tickets or cancellations. Cross-disciplinary effort is needed to even define such metrics, as understanding from modeling, engineering and business angles is required. Once defined, these metrics are used for monitoring of the production environment and for quality control of model updates.

### 5.2 Test based verification

Test-based verification is intended for ensuring that the model generalizes well to the previously unseen data. While collecting validation dataset is usually not a problem, as it can be derived from splitting the training dataset, it may not be enough for production deployment.

Full scale testing in real-world environment can be challenging for a variety of safety, security and scale reasons, and is often substituted with testing in simulation. That is the case for models for autonomous vehicles control [12]. Simulations are cheaper, faster to run, and provide flexibility to create situations rarely encountered in real life. Thanks to these advantages, simulations are becoming prevalent in this field. However, it is important to remember that simulation-based testing hinges on assumptions made by simulation developers, and therefore cannot be considered a full replacement for real-world testing. Even small variations between simulation and real world can have drastic effects on the system behavior, and therefore the authors conclude that validation of the model and simulation environment alone is not enough for autonomous vehicles.

# 6. INTEGRATION

The model integration step constitutes of two main activities: building the infrastructure to run the model and implementing the model itself in a form that can be consumed and supported. While the former is a topic that belongs almost entirely in systems engineering and therefore lies out of scope of this work, the latter is of interest for our study, as it exposes important aspects at the intersection of ML and software engineering. In fact, many concepts that are routinely used in software engineering are now being reinvented in the ML context.

# 7. MONITORING

Monitoring of evolving input data, prediction bias and overall performance of ML models is an open problem. Another maintenance issue highlighted by this paper that is specific to data-driven decision making is feedback loops. ML models in production can influence their own behavior over time via regular retraining. While making sure the model stays up to date, it is possible to create feedback loop where the input to the model is being adjusted to influence its behavior. This can be done intentionally, as well as happen inadvertently which is a unique challenge when running live ML systems.

# 8. UPDATING

Once the initial deployment of the model is completed, it is often necessary to be able to update the model later on in order to make sure it always reflects the most recent trends in data and the environment. There are multiple techniques for adapting models to a new data, including scheduled regular retraining and continual learning . Nevertheless in production setting model updating is also affected by practical considerations.

# 9. OTHER ASPECTS

## 9.1 Ethics

Ethical considerations should always inform data collection activities. As stated in the report on ethical AI produced by the Alan Turing Institute [20], “it is essential to establish a continuous chain of human responsibility across the whole AI project delivery workflow”. If researchers and developers do not follow this recommendation, complications may come up due to a variety of reasons: breach of governmental regulations, unjustifiable outcomes, aggravation of existing issues, and more [20]. Various countries have produced regulations to protect personal data rights.

## *9.2* End users’ trust

To overcome this skepticism, the team prioritized building trust, by:
Building strong communication channels;

- Sharing with stakeholders progress towards developing the goal instead of showing technical advances;
- Establishing mechanisms for public and external accountability;
- Engaging both front-line clinicians and enterprise-level decision makers at early stages of the project.

## 9.3Security

We focus specifically on adversarial machine learning and consider other related general security concerns in deploying systems such as access control and code vulnerabilities beyond the scope of our work. In data poisoning, the goal of the adversarial attack is to deliberately corrupt the integrity of the model during the training phase in order to manipulate the produced results. Poisoning attacks are particularly relevant in situations where the machine learning model is continuously updated with new incoming training data. Jagielski et al. reported that in a medical setting using a linear model, the introduction of specific malicious samples with a 8% poisoning rate in the training set resulted in incorrect dosage for half of the patients [23].

# 10. POTENTIAL SOLUTIONS

This survey looked at case studies from a variety of industries: computer networks, manufacturing, space exploration, law enforcement, banking, and more. However, further growths of ML adoption can be severely hindered by poor deployment experience. To make the ML deployment scalable and accessible to every business that may benefit from it, it is important to understand the most critical pain points and to provide tools, services and best practices that address those points. The market for machine learning tools and services is experiencing rapid growth As a result, tools for individual deployment problems are continuously developed and released. For some of the problems we have highlighted, making use of the right specific tool is an entirely reasonable approach.

.

# 11. CONCLUSION

In this survey, we showed that practitioners deal with challenges at every step of the ML deployment workflow due to practical considerations of deploying ML in production. We discussed challenges that arise during the data management, model learning, model verification and model deployment stages, as well as considerations that affect the whole deployment pipeline including ethics, end users' trust and security.

We hope this survey will encourage discussions within the academic community about pragmatic approaches to deploying ML in production.

# References

[1] Arif Cam, Michael Chui, and Bryce Hall. Global AI survey: AI proves its worth, but few scale impact. McKinsey Analytics, 2019.

[2] Thomas H. Davenport and Rajeev Ronanki. Artificial intelligence for the real world. Harvard business review, 96(1):108–116, 2018.

[3] Royal Society (Great Britain). Machine Learning: The Power and Promise of Computers that Learn by Example: an Introduction. Royal Society, 2017.

[4] Michal Pechou ˇ cek and Vladimír Ma ˇ ˇrík. Industrial deployment of multi-agent technologies: review and selected case studies. Autonomous agents and multi-agent systems, 17(3):397– 431, 2008.

[5] Kyle Wiggers. Algorithmia: 50% of companies spend between 8 and 90 days deploying a single AI model, 2019. Available at https://venturebeat.com/2019/12/11/algorithmia-50-of-companies-spend-upwards-of-three-months-deploying-a-single-ai-model/.

[6] Lawrence E. Hecht. Add it up: How long does a machine learning deployment take?, 2019. Available at https://thenewstack.io/add-it-up-how-long-does-a-machine-learning-deployment-take/.

[7] Kyle Wiggers. IDC: For 1 in 4 companies, half of all AI projects fail, 2019. Available at https://venturebeat.com/2019/07/08/idc-for-1-in-4-companies-half-of-all-ai-projects-fail/.

[8] Ben Lorica and Nathan Paco. The State of Machine Learning Adoption in the Enterprise. O'Reilly Media, 2018.

[9] The state of development and operations of AI applications. Dotscience, 2019. Available at https://dotscience.com/assets/downloads/Dotscience_Survey-Report-2019.pdf.

[10] Artificial intelligence and machine learning projects are obstructed by data issues. dimensional research, 2019. Available at https://telecomreseller.com/wp-content/uploads/2019/05/EMBARGOED-UNTIL-800-AM-ET-0523-Dimen.

[11] Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. arXiv preprint arXiv:1905.04223, 2019.

[12] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control, 2019.

[13] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics, 18(12):2917–2926, 2012

[14] Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training NLP models: A concise overview. arXiv preprint arXiv:2004.08900, 2020.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[16] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415:295–316, 2020.

[17] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1):6765–6816, 2017.

[18] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In Advances in neural information processing systems, pages 2951–2959, 2012

[19] Lucas Bernardi, Themistoklis Mavridis, and Pablo Estevez. 150 successful machine learning models: 6 lessons learned at Booking.com. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1743–1751, 2019.

[20] David Leslie. Understanding artificial intelligence ethics and safety. arXiv preprint arXiv:1906.05684, 2019.

[21] Nantheera Anantrasirichai and David Bull. Artificial intelligence in the creative industries: A review, 2020.

[22] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. arXiv preprint arXiv:2004.11138, 2020.

[23] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP), pages 19–35. IEEE, 2018.

[24] Neil Lawrence. Modern data oriented programming, 2019. Available at http://inverseprobability.com/talks/notes/modern-data-oriented-programming.html.

[25] Tom Borchert. Milan: An evolution of data-oriented programming, 2020. Available at https://tborchertblog.wordpress.com/2020/02/13/28/.

[26] Alexander Lavin and Gregory Renard. Technology readiness levels for machine learning systems. arXiv preprint arXiv:2006.12497, 2020.