TOWARDS EXPLAINABLE MACHINE LEARNING OPERATIONS (MLOPS)

Krzysztof Dzieciolowski and Ningsheng Zhao
Concordia University / Daesys. Montreal, Quebec, Canada

ABSTRACT

In the dynamically expanding AI applications, there is a need for an effective operational framework for managing ever increasing number of predictive models. A new field of Machine Learning Operations (MLOps) has recently emerged to address the issue of efficient operations and management of machine learning models. However, enterprises still struggle to develop effective MLOps systems due to lack of expertise and limited experience in AI operational needs. Attempts of an ad-hoc administration of ML algorithms have not been successful to deal with the challenges of fast proliferation of AI models. At the same time, existing concepts of MLOps frameworks don't adequately address an important question of accurate machine learning models diagnostics and explainability. In this reflection paper, we review commonly used MLOps approaches and explainability methods and suggest novel methods to address challenges of current model explainability methods. We discuss the need for university curriculum to address the issues facing management of AI models, their diagnostics and explainability.

KEYWORDS

Machine Learning Operations (MLOps), Model Explanations, Model Diagnostics, Importance Sampling

1. INTRODUCTION

Managing production of machine learning models has become challenging even for large organizations. To assist with this task a new field of Machine Learning Operations (MLOps) has emerged defined as a collaborative area of data scientists, data engineers and IT specialists. In most business and scientific applications, machine learning models are utilized repeatedly over time, to satisfy requirements for continuous usage of models' predictions. For example, the marketing department of a financial company may want to utilize machine learning models' predictions in its monthly campaigns to identify customers who are most likely to accept their credit card offer. These recurring predictions need to be updated, verified, and explained periodically with the ever-changing customer data. Given the need of recurring campaigns, there has been a requirement to maintain quality of models' predictions through automation of their updates, maintenance, diagnostics, and explanations.

A broad introduction to Machine Learning Engineering (MLE) of which MLOps is an integral part has been given by Burkov (2020). The author has distinguished nine principal stages of the machine learning project lifecycle: 1) goal definition, 2) data collection and preparation, 3) feature engineering, 4) model training, 5) model evaluation, 6) model deployment, 7) model serving, 8) model monitoring, and 9) model maintenance as shown in Figure 1. The project lifecycle stages 1) through 4) are concerned with data science model development process, while stages 6) through 9) are related to IT driven models' management and operations that can be construed as MLOps. However, such a definition of MLOps suffers from the narrow data engineering perspective and fails to recognize importance of data science needs for continuous machine learning models operations, diagnostics and explainability. Symeonidis et al (2022) provides a review of to-date, limited literature of the new MLOps field and suggests use of AutoML (Automated Machine Learning) retraining within MLOps framework so to obtain higher degree of models' accuracy. He also suggests that for high maturity MLOps systems, there is a need to complement models' monitoring with a capability for robust models explainability, though no further details are provided.

This paper has two main objectives: first, to provide suggestions for a modified design of a MLOps framework that includes capability to operate, accurately diagnose and explain ML models, and second, suggest new methods of explaining machine learning models that do not rely on an unrealistic assumption of

feature independence. These two objectives are complementary and lead towards comprehensive and explainable MLOps architecture.

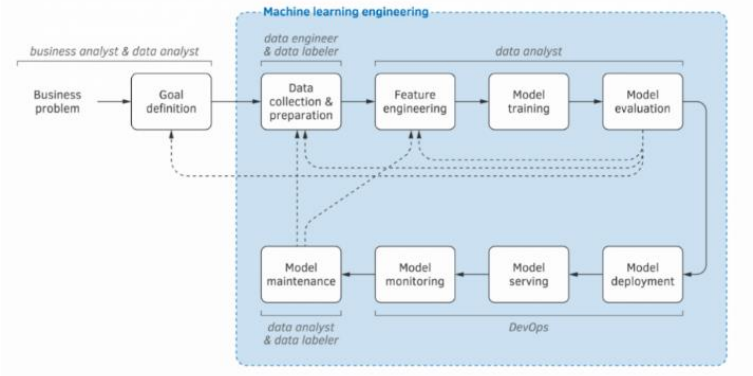

Figure 1. Machine Learning project life cycle

In section 2 challenges facing existing MLOps frameworks are reviewed, and enhancements are suggested. In section 3, a brief review of challenges of machine learning model explanations are provided and a novel approach to remedy untenable common assumption of uncorrelated features is provided. In section 4, we provide suggestions for further research in the area MLOps as well as model diagnostics and explainability.

2. IMPROVEMENTS TO MLOPS SYSTEMS

A growing need for MLOps systems can be illustrated by the fact that according to Forbes (2019) only 10% of machine learning models have been deployed into production. In addition, a single machine learning model deployment may take up to 3 months (VentureBeat, 2019). The growing use of AI is a challenge for most companies (Tech Republic, 2019). Traditionally MLOps systems have narrowly focused on solving the operational IT issues. On the other hand, we believe that the data science perspective of machine learning models diagnostics and explainability have not been so-far adequately addressed. We suggest that MLOps systems should contain an internal, robust capability for model quality monitoring and features explanations.

2.1 Classifier Rank

Typical machine learning model monitoring measures include metrics derived from the confusion matrix, such as F1, Receiver Operating Curve (ROC), Lift, Gain and Precision-Recall Curve (PRC) and others. However, these measures are sensitive to the imbalance rate, proportion of positive events. To remedy this problem, Zhao et al. (2022) suggested a new, robust measure of Classified Rank (CR) which provides the rank of the classifier in the space of all possible classifiers. Classifier Rank measure should be used for robust monitoring of models' quality in MLOps systems.

2.2 Model Explanations

Explaining machine learning models, often deemed as black box models, is an important practical problem.

Shapley values that have been derived from a game theoretic concept can be used to explain individual predictions and provide interpretation to a machine learning model. Aas et al. (2019) noted that Shapley value framework has desirable theoretical properties and can in principle handle any predictive model. Ribeiro et al. (2016) pointed out a question of trust (confidence) in a model. Therefore, monitoring explanations over time may reveal a potentially changing relationship between model predictions and features. MLOps systems that have capability to monitor model explanations may provide insights into their evolution. Both global and local model explanations should be used depending on the purpose of the model explanations. Fast developing applications of AI demand that universities adapt their curricula to prepare students for emerging technologies such MLOps systems and address interpretability of machine learning models to enable their more responsible applications.

3. A NEW APPROACH TO MODEL EXPLANATIONS

A popular way to explain machine learning models is understanding which features are important to the model output. Many possible approaches, such as L2X, LIME, RISE, etc., have been proposed to assign important scores to features, but Lundberg et al. (2017) shows that Shapley value attribution is the unique solution that satisfies certain desirable properties.

3.1 Shapley Value Explanation

Consider a machine learning model f that takes a set of features, denoted as $D = \{1,2,\dots,d\}$, as input and generates a prediction as output. We use the lowercase symbol $x = (x_1,\dots,x_d)$ to denote the values of the input vector and the uppercase symbol $X = (X_1,\dots,X_d)$ to denote the corresponding random variables measuring the input vector. Then, the Shapley value attribution of feature i can be calculated as

$$\phi_i(v) = \sum_{S \in D\setminus\{i\}} \frac{|S|!\,(|D|-|S|-1)!}{|D|!}\big(v(S\cup\{i\}) - v(S)\big),$$

where $v(S)\colon 2^d \to \mathbb{R}$ is a value function representing the model's output when only features in subset $S \subseteq D$ are considered. The Shapley value attribution captures the average marginal contribution, $v(S\cup\{i\}) - v(S)$, of feature i across all possible subsets of features that exclude i. To compute it, we must first specify the value function $v(S)$. However, the machine learning model f requires the input values $x = (x_1,\dots,x_d)$ of all features for prediction, so we cannot just remove features in $\bar{S} = D \setminus S$ and ask f to make prediction by only inputting a sub-vector x_S. There are two popular solutions: off-manifold and on-manifold value functions.

3.2 Proposed Importance Sampling Approach

To incorporate feature dependence into the Shapley value explanation, we propose the importance sampling method to approximate the on-manifold value function $v(S)$. Specifically, given $X_S = x_S$, we first draw a sample set $\{x_{\bar{S}}^{(1)}, x_{\bar{S}}^{(2)},\dots,x_{\bar{S}}^{(K)}\}$ from a known and easily sampled distribution $q(X_{\bar{S}}|X_S = x_S)$. Then, we can approximate $v(S)$ using the weighted average of $f\big(x_S, x_{\bar{S}}^{(k)}\big), k = 1,\dots,K$, over all samples, i.e.,

$$v(S) \approx \frac{1}{\sum_{k=1}^{K} w^{(k)}} \sum_{k=1}^{K} w^{(k)} f\big(x_S, x_{\bar{S}}^{(k)}\big), \text{ where } w^{(k)} = \frac{p\big(X=(x_S, x_{\bar{S}}^{(k)})\big)}{q\big(X_{\bar{S}} = x_{\bar{S}} \big| X_S = x_S^{(k)}\big)}.$$

We refer to $w^{(k)}$ as the *importance weight* of each sample $x_{\bar{S}}^{(k)}$, and trivial importance weights will be assigned to "impossible samples" with $p\big(X = (x_S, x_{\bar{S}}^{(k)})\big) \approx 0$, especially when it shows a high density value in distribution $q(X_{\bar{S}}|X_S = x_S)$. This helps prevent $v(S)$ from being estimated off the true data manifold, and thereby producing more informative model explanations. To compute $w^{(k)}$, we need to (1) train a density estimator $\hat{p}(X)$ to learn the true distribution $p(X)$, and (2) select the distribution $q(X_{\bar{S}}|X_S = x_S)$.

Training a density estimator $\hat{p}(X)$ is a common unsupervised learning problem. We need to choose and fit a state-of-the-art neural density estimator on the training or testing data. However, the selection of $q(X_{\bar{S}}|X_S = x_S)$ is a challenge, because it needs to be: (1) easy for sampling given $X_S = x_S$ for an arbitrary subset S; (2) fast to compute the conditional density value for samples; (3) as close to $p(X_{\bar{S}}|X_S = x_S)$ as possible to reduce the proportion of "impossible samples". In this work, we propose using a conditional mixture model for sampling. This approach assumes that the underlying distribution of interest contains J components, and within each component, all features are independent. In other words, for each component j, each feature i has its own marginal distribution $q_{ij}(X_i)$, such as Gaussian distribution. We further introduce a J-dimension multinomial random variable $Z \sim$ *multinomial(1, π)* as the indicator of component, where $\pi = (\pi_1,\dots,\pi_J)$ with each $\pi_j = P(Z_j = 1)$, which is referred to as the *component weights*. Then, the density function of the whole distribution can be written as $q(X) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} q_{ji}(X_i)$, and E-M algorithm can be used to estimate the unknown parameters π and each $q_{ij}(\cdot)$. Once the above mixture model is learned, given any subset of features $X_S = x_S$, we can easily compute the conditional distribution,

$$q(X_{\bar{S}}|X_S = x_S) = \sum_{j=1}^{J} P\big(Z_j = 1\big|X_S = x_S\big) \prod_{i\in\bar{S}} q_{ji}(X_i),$$

which can be referred to as the *conditional mixture model*, and where the posterior probability is given by,

$$q\big(Z_j = 1\big|X_S = x_S\big) = \frac{\pi_j \prod_{i\in S} q_{ji}(X_i = x_i)}{\sum_{j=1}^{J} \pi_j \prod_{i\in S} q_{ji}(X_i = x_i)}.$$

This conditional mixture model can be used for sampling as the following process: 1. Sample a $z \sim q\big(Z_j = 1\big|X_S = x_S\big)$, and return the corresponding j such that $z_j = 1$; 2. For each $i \in \bar{S}$, sample $x_i \sim q_{ij}(X_i)$.

The proposed importance sampling method has the following new contributions:

- It requires fitting two models $\hat{p}(X)$ and $q(X)$ only over the N samples of training data, rather than over $N \times 2^d$ samples, which are much easier to train and evaluate.
- It can well keep the estimation on the data manifold via a state-of-the-art neural density estimator $\hat{p}(X)$, such as Roundtrip, RealNVP and MAF.
- It can mitigate the data sparsity and the curse of dimensionality by re-sampling from a well-trained conditional mixture model.
- It is scalable to high dimension d, with computational complexity $O(KJd)$.

4. CONCLUSION

We have observed that current MLOps systems do not provide flexibility of managing, diagnosing, and explaining machine learning models. An expanded model diagnostics and explanation measures are needed to better understand model performance and its interpretability over time. A novel Classified Rank measure that considers data imbalance rate can be used to provide more robust and accurate assessment of model performance over time. We suggested a framework of importance sampling to overcome an unrealistic assumption of features' independence when explaining models. We point out the need to help university students understand the challenges facing organizations scaling adoption of machine learning models.

ACKNOWLEDGEMENT

The first author would like to acknowledge financial support from Concordia University Part-time Faculty Association CUPFA. The second author wants to acknowledge financial support from MITACS and Daesys.

REFERENCES

Burkov, Andriy., 2020. *Machine Learning Engineering.* True Positive. ISBN-13 978-1999579579.

Dinh, L., et al., 2016. Density estimation using real nvp. *ArXiv,* abs/1605.08803.

Lundberg, S. and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems.* California, USA, pp. 4768–4777.

Masayoshi, M. et al., 2019. Explaining black box decisions by shapley cohort refinement. *ArXiv,* abs/1911.00467.

Papamakarios, G., et al., 2017. Masked autoregressive flow for density estimation. *ANIPS*, Vol. 30.

Ribeiro, M. T., et al. (2016). Why should I trust you? Explaining predictions of any classifer. *Proceeedings of te 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM.*

Sundararajan, M. and Najmi, A. 2020. The many Shapley values for model explanation. *Proceedings of the 37th International Conference on Machine Learning*. Virtual Event, pp. 9269–9278.

Symeonidis, A., et al. Definitions, Tools and Challenges. *IEEE Annual Computing and Communication Workshop and Conference (CCWC).*

Zhao, N., Yu, J.Y., Dzieciolowski, K. 2022. Classifier Rank – A new Classification Assesment Method. *7th International Conference on Big Data Analytics, Data Mining and Computational Intelligence (BIGDACI)*. Lisbon.