Disseminating Machine Learning to Domain Experts: Understanding Challenges and Opportunities in Supporting a Model Building Process

Ray (Sungsoo) Hong
ViDA, New York University
New York City, United States
rayhong@nyu.edu

Jorge Piazentin Ono
ViDA, New York University
New York City, United States
jorgehpo@nyu.edu

Juliana Freire
ViDA, New York University
New York City, United States
juliana.freire@nyu.edu

Enrico Bertini
ViDA, New York University
New York City, United States
enrico.bertini@nyu.edu

ABSTRACT

Machine Learning (ML) is increasingly becoming an available toolkit for domain experts who work in a wide range of application areas. However, for the domain experts who don't have profound knowledge about ML, building a ML model that can accurately solve their problem can be challenging. To help domain experts' model building process, researchers have put effort into building contributions in interactive machine learning and/or devising a way to improve interpretability of ML models. Nevertheless, recent findings imply that it is challenging to identify practical design guidelines that inform researchers and practitioners on how they can effectively support the domain experts' model

CHI'19, April 2019, Glasgow, UK

building process. In this paper, we aim to build a system that can effectively facilitate a model building process of domain experts who often have little or no knowledge of ML. As the first step to build such a system, we built an initial prototype with the goal of establishing a deeper understanding in terms of how domain experts go through the model building process, what are the common bottlenecks they encounter and which strategies they apply to overcome them. We think a user's model building as the learning process by which the experts actively seek new information and integrate the new with their existing knowledge. Based on our viewpoint, we believe that, by observing how the patterns of the common bottlenecks and the strategies change based on domain experts' degree of ML-related knowledge and usage context (e.g., in-the-lab vs. in-the-wild), we will gain useful insights into the model building process. We hope the insights can be synthesized with the existing findings and contribute to formalizing design guidelines that can present practical insights to those who aim to support domain experts' model building process.

KEYWORDS

Human-centered machine learning, applied machine learning, model building process

INTRODUCTION

Advances in machine learning (ML) have demonstrated that it is possible to train models that outperform humans on a variety of tasks (e.g., AlphaZero, Google PlaNet-photo [13]). As ML has become a powerful solution, studies in applied ML, Human-Computer Interaction (HCI), and Information Visualization (InfoVis) have focused on helping data scientists more easily interact with ML-related systems. For instance, researchers have studied implications, techniques, systems related to ML interpretability[3, 7] to assist data scientists to understand why a certain ML model made a specific prediction under a given circumstance[3]. While the existing research focuses on understanding effectively support data scientists who generally have an intermediate or expert level of ML knowledge[14], the user base of ML is growing even further, ranging from data scientists to domain experts who often have little or no ML knowledge[7]. However, understanding the effective design that enables a novice user's model building process is a profound research challenge. To disseminate ML to a broader user base, we see it is essential to gain insights that can guide designing a system that can effectively support domain experts' model building process who may have different ML-related knowledge.

In this work, we aim to develop a system that domain experts can easily interact with for building their predictive model. Within the past years, we have developed a system named NYU-TA3, under the support from DARPA D3M program [1]. Our system synthesizes end-to-end pipelines with a user in the loop which enables a workflow of a model building process that includes loading and augmenting datasets, defining a problem (i.e., setting features and outcome), creating possible models, assessing the validity of the models, and deployment. The overarching challenges we encountered at this point

[1] More information about Data-Driven Discovery of Models (D3M) can be found at: https://www.darpa.mil/program/data-driven-discovery-of-models

is that there are no design insights or guidelines that inform us to understand how to effectively support domain experts' model building process. We see this challenge as significant because (1) the workflow itself includes a variety of user stages that each requires different types of ML-related knowledge, while (2) domain experts' ML-related knowledge and their prior domain knowledge may vary. In sum, it is difficult to understand the gap between the knowledge that domain experts need to understand to build a model and the knowledge that they actually have. As our initial investigation to deepen our understanding of a better design for supporting the domain expert's model building process, we aim to use TA3 as a design probe to uncover the perspectives as follows: (1) the common bottlenecks that domain experts encounter in each stage and the reasons behind, (2) the common ways that the domain experts apply to deal with when they encounter the bottleneck, and (3) the common usage patterns of a system that show which features the domain experts use for improving their model. We expect having an understanding of the above perspectives may help future designers to understand the common knowledge gap that users encounter in each stage and how to use existing techniques to resolve the gap.

We expect that our further investigation in the directions we identified above can offer contributions to the domain of Human-Centered Machine Learning as follows. First, having a deeper understanding of the common bottlenecks that domain experts encounter in each stage of their model building process and the common strategies they apply may present useful insight for implementing easy-to-use tools and broadening user base of ML. Second, iterating our system based on the insights we found may enable building a unique artifact that improves the current practice of model building support tools. Finally, the improved NYU-TA3 may be used as a test-bed for observing how even a broader range of in-the-wild domain experts would build their ML model. Such observation may discover new research opportunities.

BACKGROUND

Designing interactive systems that can support a user's model building process has been studied in many studies. Perhaps one of the earliest approach that presents graphics to help a user's model building process is Ware et al.'s work[12]. Later, Fails and Olsen have suggested the fashion of interactive ML where a user takes an active role in training and correcting a model [4]. Based on the approaches, more applied visualization techniques and interactive systems have been suggested in HCI and InfoVis communities. For instance, EnsembleMatrix [11] suggests a technique that combines a set of models to construct one's own. FeatureInsight [2] and ModelTracker [1] each presents design of an analytic system for improving a feature ideation and model building process. Squares apply visualization techniques to support performance analysis of multiple models [9]. In recent years, there has been growing research interest in understanding the implication of interpretability in ML models [3], as well as some specific techniques that give an explanation to a user in terms of why a model made a prediction in a certain way. For instance, Anchor is a model-agnostic system that explains a given model's instance predictions [10], while Manifold [15] presents an interactive system that allows a user to interpret, debug, and compare multiple models. RuleMatrix help a user to understand, explore and validate predictive models.[7]

The aforementioned techniques and systems present a set of novel features that empower a user who attempt to improve their model. In general, these approaches are built based on the assumption that a user of their techniques or systems is an advanced statistician or a research staff that has a certain knowledge related to ML. Meanwhile, as ML performance is improving and infrastructure required for building model is becoming cost-efficient, ML is becoming an available toolkit to a more general audience [7]. AutoML tools such as Google AutoML [5], and Microsoft Machine Learning Studio[6] enable general users to automatically generate ML pipelines tailored for their datasets. Despite the recent trend, the model building process is known to be challenging for a user who has little or no ML related knowledge. There is relatively scarce of studies dedicated to understanding how to effectively support novice ML users' model building process (e.g., TPOT[8], RuleMatrix [7]).

Recent studies started investigating how "Non-experts" use tools to build ML models. The focus of their study is to understand people who don't have knowledge about ML, but the focus is for people who already started using ML building tools, often with the support from ML consultants (e.g.,[14]). We think deepening our understanding of the struggles that the *novice* domain experts would encounter when they first experience the tools and their learning process about ML knowledge would help us to identify more generalizable findings for benefiting a broader range of users.

METHOD

We briefly explain TA3 then discuss our considerations in terms of collecting data from participants.

TA3 is an interactive system that allows users to create predictive models. It presents a step-by-step user interface which enables users can follow to build their model. In particular, the specific user stage of the system is presented as follows:

- **Prepare data**: A user can search available datasets. A user can also load one's own datasets. Once a data is loaded, one can augment the data (i.e., search relevant datasets and join some features).
- **Define a problem**: A user can set a series of items required for model creation. For example, one can select data features to use, a problem type (e.g., classification, regression, clustering).
- **Create models**: A user can specify how much computational resources they will allocate for creating models.
- **Examine models**: A user can assess and compare multiple models using charts and graphs, and decide which model to apply.

Each user stage has multiple interactive features a user can leverage and we expect a user needs to know or learn a certain level of ML related knowledge to move forward. We introduce the two analysis points (A1, A2, independent variables) and the two observation points (O1, O2, dependent variables) as follows:

- **A1. A degree of prior knowledge**: One factor we think extremely critical is to measure the degree to which our participants know about ML. More specifically, our primary expectation is that based on the fact as to whether they had some ML-related training or not, the patterns we may observe in A1 and A2 may vary. Examining the above two based on their degree of ML knowledge may present some useful insights to the HCI community.
- **A2. The context of use, in the lab vs in the wild**: Our grand mission behind NYU-TA3 is to disseminate ML to broader audiences and benefit people by enabling them to make data-driven decision-making. Therefore, another perspective we are particularly interested is to see how the domain experts working in the real world may use TA3. We are expecting to run the study in-the-lab, but also in-the-wild to capture real context which may leave us some research challenges.
- **O1. Understand the common bottleneck**: Our core interest is to understand the common bottleneck that may befuddle users while they interact with multiple features in each user stage. We are particularly interested to know whether the bottleneck happened because of people's lack of knowledge related to ML, and if presenting some relevant material that can educate them would enable them to understand the meaning of features. We are also interested in understanding as to whether what are the features they feel necessary and useful, what features are advanced, or what features are needed once a user fully understood the functionality. We hope having a comprehensive understanding of these steps can yield some general insights that the HCI community can share.
- **O2. Understand people's patterns for building their model**: It wouldn't be surprising if we imagine the model building process can be highly explorative and iterative. We wish to track user stages that each user visit for accomplishing their task (i.e., finish building a model) and see if their way of visiting stages shows linear or jumping back-and-forth. If we identify jumping, we would be curious to know why a user jumped to some another stage for which reason. If we identify something interesting, we would like to establish some model that can explain a user's model building process.

CONCLUSION

We assume the patterns we can observe in terms of O1 and O2 may vary based on the condition of A1 and A2. With a deeper understanding of why such different patterns have been observed, we may be able to draw useful design insights that could evolve to a more generalizable guidelines for supporting domain experts' model building process.

ACKNOWLEDGEMENTS

This project was partially supported by the NSF, the Moore-Sloan Data Science Environment, the DARPA D3M program and Labex DigiCosme. Any opinions, findings, and conclusions or recommendations in this material are those of the authors and don't necessarily reflect the views of DARPA.

REFERENCES

[1] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 337–346.

[2] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE, 105–112.

[3] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[4] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.

[5] Google. 2019. Google Cloud AutoML. https://cloud.google.com/

[6] Microsoft. 2019. Microsoft Machine Learning Studio. https://studio.azureml.net/

[7] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE transactions on visualization and computer graphics* 25, 1 (2019), 342–352.

[8] Randal S Olson and Jason H Moore. 2016. TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on Automatic Machine Learning*. 66–74.

[9] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2017. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 61–70.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.

[11] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. 2009. EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1283–1292.

[12] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. 2001. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies* 55, 3 (2001), 281–292.

[13] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*. Springer, 37–55.

[14] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*. ACM, 573–584.

[15] et al. Zhang. 2019. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE transactions on visualization and computer graphics* 25, 1 (2019), 364–373.