On the Interaction between Software Engineers and Data Scientists when Building Machine Learning-Enabled Systems

Gabriel Busquim, Maria Julia Lima (coorientadora), Marcos Kalinowski (orientador)

Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, RJ – Brazil

{gbusquim, kalinowski}@inf.puc-rio.br, mjulia@tecgraf.puc-rio.br

Abstract. *Engineering ML-enabled systems presents various challenges from both a theoretical and practical perspective. One of the key challenges is the effective interaction between actors with different backgrounds who need to work closely together, such as software engineers and data scientists. This dissertation involved three studies investigating the current collaboration dynamics between these two roles in ML projects. Our studies revealed several challenges that can hinder collaboration between software engineers and data scientists, including differences in technical expertise and unclear definitions of each role's duties. Potential solutions to address these challenges include encouraging team communication and producing concise system documentation.*

1. Background and Motivation

Integrating ML components into existing systems has increased as companies seek to leverage large amounts of data to enhance the business outcomes of their software products. We refer to these systems as ML-enabled systems, since their behavior is dictated by explicitly defined rules and the data used by the ML component to make decisions. This transition from traditional software systems to those integrated with ML components introduces new challenges from the perspective of software engineering [Nahar et al. 2022]: designing an appropriate architecture for these systems is not trivial, and the different backgrounds of data scientists and software engineers may cause barriers in a collaborative environment. Our goal with this dissertation was to thoroughly investigate how software engineers and data scientists collaborate when developing ML-enabled systems. In addition, we sought to obtain recommendations to promote better collaboration practices and uncover future work possibilities for other researchers interested in this field.

2. Methodology

To accurately characterize collaboration in a real-world context, we discussed this topic with data scientists and software engineers currently developing ML-enabled systems for industry-academia projects. We conducted three distinct studies, each involving different teams and participants. The first one consists of a case study with a team developing an ML-enabled system for legal conflicts resolution [Busquim et al. 2024b], where we interviewed the team's software engineers and data scientists to uncover their collaboration practices. In our second study, we interviewed members of two other teams building ML-enabled systems for a customer in the Oil and Gas sector. Finally, our third study comprised focus groups with experienced data scientists and software engineers

[Busquim et al. 2024a]. This study examined how collaboration can be relevant during various development tasks, as well as the effectiveness of several recommendations we proposed to improve this interaction based on our findings and the current literature.

3. Results and Concluding Remarks

Our results provide factual examples of collaboration challenges based on real ML-enabled system projects for different customers. The challenges we described are also stated in the current literature, which confirms their relevance. In our first and second studies, participants revealed the obstacles they faced due to ineffective collaboration during tasks such as integrating the ML model with other system components and updating project documentation. This allowed us to propose several recommendations to prevent this scenario, such as establishing clear system requirements and fostering a collaborative environment. As a result of our third study, we obtained a detailed view of the importance of collaboration during multiple technical tasks vital for ML-enabled system development. We were also able to assess the importance of each recommendation we had defined for evaluation, and how teams can use them to improve their performance.

Overall, our research contributes to understanding the complex dynamics between software engineers and data scientists in ML projects and provides insights to improve collaboration and communication in this context. We encourage future studies investigating this interaction in other projects, as this can both validate and enrich our findings.

References

Busquim, G., Araújo, A., Lima, M., and Kalinowski, M. (2024a). Towards effective collaboration between software engineers and data scientists developing machine learning-enabled systems. In *Anais do XXXVIII Simpósio Brasileiro de Engenharia de Software*, pages 24–34, Porto Alegre, RS, Brasil. SBC.

Busquim, G., Villamizar, H., Lima, M. J., and Kalinowski, M. (2024b). On the interaction between software engineers and data scientists when building machine learning-enabled systems. In *International Conference on Software Quality*, pages 55–75. Springer.

Nahar, N., Zhou, S., Lewis, G., and Kästner, C. (2022). Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th international conference on software engineering*, pages 413–425.