Business Case Summary for Implementing MLOps at Kronkers

By: Kevin Sandoval

A.

MLOps is “the process of managing the machine learning life cycle from development to deployment and monitoring “ (Sewell, D602 Webinar # 2). In his webinar, Dr. Sewell showcases that the process begins with choosing a proper objective and knowing what the desired results, processes, and failures can be. After the objective is clear and determined, the process is then kicked off with the data engineering side gathering and storing the data necessary to answer the objective. It is then the data scientists’ job to model the data to determine if there are discernible improvements that can be made within the objective. The process is then handed over to the software engineering team to start the product development process of designing and testing integration. Finally, the process is given to the ML engineers who can deploy the new machine learning model and monitor its effectiveness and change as necessary. The process is like that of DevOps, which is “the set of practices which help deliver applications and services faster and more reliably. It can help entities respond faster to customer needs, become more competitive, increase automation, job satisfaction, security and change management” (Sewell, D602 Webinar # 2).

The main objective of MLOps is to find ways to increase profits (Huyen, 2022). This can either be seen in increasing sales or by cutting costs. Finding an effective MLOps solution can influence business performance positively as it will create a more optimal solution to company problems. It can be seen clearly in performance metrics such as amount of ads revenue, number of monthly active users, and more. It can also lead to better quality assurance, faster experimentation and development, and faster deployment of outdated models (Huyen, 2022).

B.

There are several constraints/considerations when trying to implement an MLOps solution. The first and most pressing is the cost. The cost is felt in both initial cost to create, but also in time and effort required to create the MLOps solution. It is recommended to use a few data engineers/data scientists in the beginning stages to try and figure out if an MLOps solution is viable, before putting more people, time, and effort into the process to bring the solution to completion (Huyen, 2022).

Another consideration when choosing an MLOps solution is to ensure that the initial objective is good. What this means is ensuring that the MLOps solution meets a set of selection criteria. In Section 2 video “Selecting ML Projects” of his LinkedIn course “*MLOps Essentials: Model Development and Integration”*, Ponnambalam (2022) lists 6 important criteria.

The first is core business values. This is referring to the fact that the solution may be expensive and may not have long term value in the organization. If that is the case, then the solution does not meet the core business value of creating profits (Ponnambalam, 2022).

The second is training data availability. This is an important step at the beginning of the process for the data engineers and data scientists. If the organization is trying to use bad data to create a model, the model and the ML project will fail, resulting in a large loss in time and money (Ponnambalam, 2022).

The third criterion is the technology ecosystem. This is referring to the algorithms, libraries, framework, and pre-trained models that may already exist within the organization. Standardizing can be possible but it is time consuming and potentially more expensive, which could be the case for Kronkers (Ponnambalam, 2022).

The fourth criterion is team size and composition. He recommends starting with a smaller team to find meaningful trends before investing a larger team. However, if there are not enough people on the team, then it will take a lot longer to create and delay the time to deploy (Ponnambalam, 2022).

The fifth criterion is time to market. MLOps is a fast moving environment and models need to be created and updated quickly. A long time to market will already be delaying potential benefit for the company and will risk becoming outdated as the team is working on it (Ponnambalam, 2022).

The sixth and last criterion is the risk of failure. Ponnambalam mentions that that organization should be made aware that failure can and will happen sometimes and that the risk involved when creating an MLOps solution is present (Ponnambalam, 2022).

C.

Kronkers could benefit from implementing an MLOps solution to address their desire to “enchance various aspects of its operations, including personalized product recommendations, demand forecasting, and customer segmentation” (Kronkers, 2024). However, there does seem to be considerable cost when starting the process. As noted in the Kronkers Project Overview document, nothing in the organization is currently standardized. Most work in python but some work in R and Julia. It would take dedicated time and effort to standardize although it would help in the long run. The current machine learning models are also not widespread and deployment varies by model. This would be another improvement to be made with an MLOps solution. Kronkers also has no currently role dedicated to the maintenance of code, machine learing models, or deployment. A key aspect of implementing an MLOps solution is to ensure that the team is not understaffed. Since Kronkers has only a small budget for machine learning model maintenance, tracking, and quality control, it might be difficult to fully implement an MLOps solution from the get-go. I would recommend using the small budget allocated in the current fiscal year to do some exploratory analysis on the data that they have access to and determine if there are significant gains to be made by the models that are created (Ponnambalam, 2022). This hedges the risk of failure and if the data engineers and data scientists are able to find beneficial insights, then it is potentially worth the time and effort to standardize everything and implement a full MLOps solution.

In this proposed solution, the nonfunctional requirements are scale, security, serviceability, deployment, and the user experience (Ponnambalam, 2022). The MLOps solution should be able to scale with growth of the customer base as well as being able to be edited for when newer models or techniques arrive. Security is important to retain intellectual property and rights to the work created within the company. Serviceability is a core belief at Kronkers and should thus be retained when creating the solution. Deployment is a late stage requirement that sees the ML engineers making it active for the users. And lastly the user experience is vital. It should be easy to use and maintain (Paraphrased from Ponnambalam, 2022).

The functional requirements are the metric goals for the model and the product/service, performance goals, the operational goals, and the cost. The cost is an important consideration and is vital in what the model and MLOps solution can produce. The performance goals for the model can be seen in the accuracy score, the F1 score, type 1 and 2 errors, loss, and the AUC score. The performance goals for the product/service are tied to the customer data such as engagement, items bought, etc. The operational goals for the product are latency, requests per second, and max concurrent sessions. The operational goals for the model are latency, resource utilization, and costs. (Paraphrased from Ponnambalam, 2022).

Clearly there are improvements to made with optimization within the Kronkers organization. These improvements require time and effort and must be considered when trying to develop an MLOps solution for them. Since there is only a small budget and a relatively small team, I think it would be best to do some exploratory practices before employing a full team.

D.

altexsoft. (November 29th, 2023). *Functional and Nonfunctional Requirements: Specifications and Types.* Retrieved December 14th, 2024 From <https://www.altexsoft.com/blog/functional-and-non-functional-requirements-specification-and-types/>

Dr. Sewell, W (n.d). *D602 Webinar # 2.* Retrieved December 8th, 2024,From D602 Course Search

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4bdc60ba-a36a-4e4e-8172-b21900d735f5>

Huyen, C (June 28th, 2022). *Designing Machine Learning Systems.* Retrieved December 10th, 2024,From D602 Course Materials

Kronkers (2024), *Kronkers Project Overview.* Retrieved December 11th, 2024, From D602 Task 1 Assessment Page

Ponnambalam, K (September 16th, 2022). *MLOps Essentials: Model Development and Integration.* Retrieved December 8th, 2024,From Section 1 *Introduction to MLOps* <https://www.linkedin.com/learning/mlops-essentials-model-development-and-integration/getting-started-with-mlops?resume=false&u=2045532>

Ponnambalam, K (September 16th, 2022). *MLOps Essentials: Model Development and Integration.* Retrieved December 8th, 2024,From Section 2 *Requirements and Design* <https://www.linkedin.com/learning/mlops-essentials-model-development-and-integration/getting-started-with-mlops?resume=false&u=2045532>