Pick Model Deployment Method

I've experimented with Batch Inference, but I'm now exploring AWS SageMaker.

Why AWS SageMaker?

While Batch Inference is free, it has several drawbacks. It's limited in scalability, lacks the ability to automatically try and use the best-suited model, and has challenges with monitoring. Moreover, it lacks the widespread industry adoption that SageMaker offers. I believe SageMaker is the right choice for me because it enhances my MLOps skills. However, it can be costly if not used wisely and effectively, which is another skill I'm learning. By choosing an efficient and cost-effective model, I can apply these MLOps skills to other projects and organizations.

Additional Benefits:

- SageMaker is highly powerful due to its integration with other AWS services and platforms, simplifying the development process. It provides robust security features to protect data. SageMaker offers real-time inference capabilities, supports multiple frameworks and languages, and includes built-in models, making overall training, deployment, and monitoring easy. Additionally, it has a supportive community and cost management strategies.