Automação no processo de publicação de modelos de Ciência de Dados

Diego Nogare[1], **Rodrigo Fernandes Mello**[2], **Marco Antonio Lopes**[1]

[1]Superintendência de Advanced Analytics (SAA)
Itaú Unibanco S.A. – São Paulo, Brasil

[2]Diretoria Estratégica de Dados (DED)
Itaú Unibanco S.A. – São Paulo, Brasil

{diego.nogare,rodrigo.fernandes-mello,marco-antonio.lopes}@itau-unibanco.com.br

Abstract. *This paper introduces the difficulties faced by data science teams while managing the life cycle of a machine learning model, from business understanding until its deployment and continuous observability. The nowadays nature of decentralized teams using different technologies makes it very difficult to publish and maintain models in company's information systems. In this context, Itaú Unibanco S.A. has designed a platform to simplify the delivery and management of those models, supporting the productive implementation and guaranteeing software engineering good practices while ensuring security and architectural pattern.*

Resumo. *Neste artigo são apresentadas as dificuldades encontradas por times de ciência de dados para fazer a gestão do ciclo de vida de um modelo de aprendizado de máquina.*
A existência de times descentralizados utilizando tecnologias diferentes dificulta muito a publicação destes modelos nos sistemas informacionais da empresa. Desta forma, o Itaú Unibanco S.A. investiu esforços na construção de uma plataforma para simplificar a gestão desses modelos e facilitar sua implantação produtiva garantindo boas práticas de engenharia de software, segurança e padrões arquiteturais

1. Contexto e Problema

As corporações atuais encontram dificuldades e limitações para a construção de um ambiente capaz de integrar as etapas de projeto e desenvolvimento de software com a publicação desses artefatos em ambientes de homologação e produtivos. Esses aspectos são ainda mais agravados na presença de múltiplos times descentralizados e distribuídos em regiões distintas de um país ou do mundo. Ademais, com a existência de múltiplas origens de dados brutos (SOR – *System of Record*) utilizadas para a produção de fontes de dados especializados (SOT – *Source of Truth* – e SPEC - *Specialized*), também descentralizadas e distribuídas, as plataformas para projeto, desenvolvimento, publicação e observabilidade de artefatos de software têm se tornado ponto focal para a Engenharia de Software moderna, incluindo sua aplicabilidade em soluções de *Machine Learning* [Serban et al. 2020].

Neste contexto, a Diretoria Estratégica de Dados (DED) do Itaú Unibanco S.A. tem investido em uma plataforma de MLOps *(Machine Learning Operations)* como produto para todos os times, origens de dados e áreas de negócio e tecnologia descentralizadas e distribuídas no território brasileiro, bem como conectando suas unidades no exterior. Essa plataforma foi projetada para execução sobre ambientes de *cloud* pública, a partir de evoluções aprendidas na operação do ambiente *on-premisses*. Sobre essa *cloud* pública, adotou-se um conjunto de soluções *open-source* e *cloud native*, combinando as melhores características de ambos, em que a primeira reduz o *lock-in* dessa plataforma, enquanto a segunda acelera seu provisionamento e reduz seu *time-to-market* para as áreas de negócio e tecnologia.

As principais dificuldades encontradas foram a integração com produtos legados, construção de novos processos para a entrega de modelos de Ciência de Dados em ambientes produtivos e adaptação dos times ao novo formato de trabalho. [Washizaki et al. 2019]. Como principal contribuição, essa plataforma possibilita a entrega rápida, seguindo boas práticas de engenharia de software, segurança e arquitetura, de soluções de aprendizado de máquina integradas com o processo produtivo. Além disso, facilita a observabilidade dos modelos e dos dados para que se possa definir momentos de retreino bem como detectar desvios nas distribuições estatísticas, tanto dos dados quanto das predições, melhorando a confiabilidade para o negócio. Uma vantagem adicional se dá pela maior transparência nas etapas de *deployment* dos modelos de Ciência de Dados em ambiente produtivo graças às governanças exigidas por esses ambientes, como gestão de mudanças e obrigatoriedade de registro de *logs*.

A experiência adquirida pelo Itaú Unibanco S.A. na construção dessa plataforma pode ainda ser utilizada na forma de oferta de serviços para empresas terceiras interessadas na construção de soluções similares.

2. Solução Proposta

As áreas de negócio e tecnologia do Itaú Unibanco S.A. foram amplamente consultadas com o intuito de descobrir e mapear as principais necessidades relativas a um ambiente que sustentasse o ciclo de vida de modelos de Ciência de Dados, Inteligência Artificial, Aprendizado de Máquina e *Deep Learning*. Esses elementos formaram, e ainda são reavaliados para formar, o *backlog* da plataforma segundo as melhores práticas da disciplina de produtos digitais e metodologias ágeis. [Serban et al. 2020, Washizaki et al. 2019, Warnett and Zdun 2022].

Em resumo, esse *backlog* motivou a construção do seguinte conjunto de componentes:

i. *Feature Store*: processo que permite a democratização e reutilização de conjuntos de dados específicos, agrupados por domínios, a fim de acelerar a criação de modelos de Aprendizado de Máquina;

ii. Ambiente de Experimentação: conjunto de ferramentas para que cientistas de dados possam trabalhar em todo o ciclo de vida do desenvolvimento, segundo metodologias de mercado como o CRISP-DM;

iii. Esteiras automatizadas de CI/CD: solução automatizada para unir códigos desenvolvidos em ramificações (*branch*) do software principal, e entregar/publicar a nova

versão do software no ambiente final;

iv. Agendamento de *queries*: rotinas de processamento e transformação de dados que necessitam de execução recorrente; [Warnett and Zdun 2022]

v. Monitoramento operacional e de desempenho de modelos de Ciência de Dados: captura, monitora e suporta a auditoria de *Logs* operacionais referentes ao ambiente utilizado, além de *Logs* de desempenho que estão relacionados à qualidade entregue pelos modelos; Acompanhamento de desvio de dados e de conceitos; [Lewis et al. 2021]

vi. Ambiente de treino e retreino de modelos de Ciência de Dados: definição de estrutura e padrão de desenvolvimento para prover escalabilidade dos modelos, bem como, alertas para que os times possam reavaliar e retreinar modelos que já não atendem mais ao desempenho esperado; [Lewis et al. 2021]

vii. Ambiente de publicação de modelos de Ciência de Dados: são permitidas as publicações de modelos de processamento *batch* e seus agendamentos, bem como os processamentos online síncrono e assíncrono; [Warnett and Zdun 2022]

Todos esses componentes visam compor uma plataforma moderna, completa e dinâmica que facilite a integração entre cientistas de dados, engenheiros de *machine learning*, engenheiros de dados e engenheiros de software para que possam entregar valor para as áreas de negócio e tecnologia de forma mais rápida e produzindo resultados relevantes para os clientes do Itaú Unibanco S.A.

O que permite que todas essas ferramentas funcionem em áreas de negócio atuando sobre problemas diversos são interfaces programáticas baseadas em serviços de automação, que buscam garantir a independência de um sistema centralizador. Além disso, cada ferramenta permite o acompanhamento das etapas do ciclo de vida de um modelo, podendo, ou não, ser utilizada em conjunto com outras ferramentas da plataforma ou de terceiros. Essa flexibilidade permite desacoplamento dos artefatos de software, aumentando o poder de customização entregue pela plataforma proposta.

Apesar da independência entre produtos da plataforma, aspectos como reprodutibilidade, qualidade e segurança são garantidos com governanças centralizadas que funcionam como requisitos mínimos para a utilização de cada uma das ferramentas. Como exemplos, pode-se listar a criação de regras de segurança mínima em cada solução, regras de gerenciamento de risco para modelos e gestão de acesso aos dados.

3. Indicador de Sucesso

O desenvolvimento desta plataforma de MLOps permitiu com que os usuários utilizassem recursos gerenciados, abstraindo conhecimentos na área de infraestrutura para realizar o provisionamento das ferramentas necessárias para sua atividade. A plataforma também habilita o uso individualizado de recursos, evitando concorrência computacional e facilitando o repasse de custos para as áreas solicitantes. Foi observado um impacto positivo no uso consciente de recursos financeiros por provisionar as ferramentas sob demanda, evitando desperdícios de capacidade computacional ociosa, o que reflete em eficiência orçamentária.

Com menos de um ano da nova plataforma, mais de 10% dos profissionais do Itaú Unibanco S.A. que desenvolvem modelos já foram para ela migrados. Já são mais de

100 modelos de Ciência de Dados publicados nos primeiros meses da nova plataforma, desconsiderando aqueles que ainda não foram para produção ou que não utilizam dados produtivos.

Com relação ao desempenho, um dos cases mais impactantes apresentou uma diminuição de mais de 15x no tempo de processamento de *insights* resultados de grafos, além de permitir ajuste da capacidade computacional de forma autônoma, sem a necessidade de abertura de chamados internos.

4. Referências

Lewis, G. A., Ozkaya, I., and Xu, X. (2021). Software architecture challenges for ml systems. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 634–638. IEEE.

Serban, A., van der Blom, K., Hoos, H., and Visser, J. (2020). Adoption and effects of software engineering best practices in machine learning. In *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12.

Warnett, S. J. and Zdun, U. (2022). Architectural design decisions for the machine learning workflow. *Computer*, 55(3):40–51.

Washizaki, H., Uchida, H., Khomh, F., and Guéhéneuc, Y.-G. (2019). Studying software engineering patterns for designing machine learning systems. In *2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP)*, pages 49–495. IEEE.