O **MLOps** ou **ML Ops** é um paradigma que visa implantar e manter modelos de aprendizado de máquina em produção de forma confiável e eficaz.[1] A expressão é uma combinação de "*machine learning*" e da prática de desenvolvimento contínuo *DevOps* na área de software. Os modelos de aprendizado de máquina são testados em ambientes isolados experimentais. Quando um algoritmo está pronto para ser lançado, ocorre a prática de MLOps entre cientistas de dados, e engenheiros de DevOps e de aprendizado de máquina para efetuar a transição do algoritmo para sistemas de produção.[2] De forma similar às abordagens de DevOps ou DataOps, o MLOps busca aumentar a automação e melhorar a qualidade dos modelos de produção, ao mesmo tempo em que foca em exigências de negócio e regulatórias. Embora o MLOps tenha começado como um conjunto de melhores práticas, ele está evoluindo lentamente para uma abordagem independente para o gerenciamento do ciclo de vida de ML. O MLOps se aplica ao ciclo de vida completo - da integração com a geração de modelo (ciclo de desenvolvimento de software, integração contínua/entrega contínua), orquestração, e distribuição, para métricas de saúde, diagnóstico, governança, e negócio. De acordo com a Gartner, o MLOps é um subconjunto de ModelOps. O MLOps é focado na operacionalização de modelos de ML, enquanto que ModelOps cobre a operacionalização de todos os tipos de modelos de AI.[3]

Definição

O MLOps é um paradigma que inclui aspectos como boas práticas, conjuntos de conceitos, bem como uma cultura de desenvolvimento no que diz respeito à conceitualização, implementação, monitoramento, distribuição, e escalabilidade de produtos de aprendizado de máquina. No geral, é uma prática de engenharia que se aproveita de três disciplinas de apoio: aprendizado de máquina, engenharia de software (especialmente DevOps), e engenharia de dados. O MLOps foca no processo de colocar em produção sistemas de aprendizado de máquina fazendo a ponte entre desenvolvimento (Dev) e operações (Ops). Essencialmente, o MLOps visa facilitar a criação de produtos de aprendizado de máquina tirando vantagem dos seguintes princípios: automação de CI/CD, orquestração de fluxo, reprodutibilidade; versionamento de dados, modelos, e código; colaboração; treinamento e avaliação contínuos de ML; rastreamento e log de metadados de ML; monitoramento contínuo; e ciclos de feedback.[4]

História

Os desafios do uso continuado de aprendizado de máquina em aplicações foi destacado em um artigo de 2015.[5] O crescimento previsto em aprendizado de máquina incluía uma duplicação nos pilotos e implementações de aprendizado de máquina de 2017 a 2018, e novamente de 2018 a 2020.[6]

Os relatórios mostram que a maioria (até 88%) das iniciativas de aprendizado de máquina enfrenta dificuldades para avançar além dos estágios de teste.[4] No entanto, aquelas organizações que realmente colocaram aprendizado de máquina em produção observaram aumentos de 3-15% na margem de lucros.[7] O valor do mercado de MLOps foi estimado em $23.2 bilhões em 2019 e projeta-se que alcançará $126 bilhões até 2025 devido a rápida adoção.[8]

Arquitetura

Os sistemas de aprendizado de máquina podem ser categorizados em oito categorias diferentes: coleta de dados, tratamento de dados, engenharia de características, rotulagem de dados, projeto de modelo, treinamento e otimização de modelos, implantação de endpoint e monitoramento de endpoint. Cada etapa do ciclo de vida do aprendizado de máquina é construída em seu próprio sistema, mas elas precisam ser conectadas. Esse é o mínimo de sistemas que as empresas precisam para dimensionar o aprendizado de máquina em sua organização.

Objetivos

Há uma série de objetivos que as empresas desejam alcançar por meio de sistemas MLOps implementando aprendizado de máquina com sucesso em toda a empresa, incluindo:[9]

- Implantação e automação[10]
- Reprodutibilidade de modelos e previsões[11]
- Diagnóstico[11]
- Governança e conformidade regulatória[12]
- Escalabilidade[13]
- Colaboração[14]
- Usos comerciais[15]
- Monitoramento e gestão[16]

Uma prática padrão, como o MLOps, leva em consideração cada uma das áreas mencionadas acima, o que pode ajudar as empresas a otimizar fluxos de trabalho e evitar problemas durante a implementação.

Uma arquitetura comum de um sistema MLOps incluiria plataformas de ciência de dados onde os modelos são construídos e os mecanismos analíticos onde os cálculos são realizados, juntamente com a ferramenta MLOps orquestrando a movimentação de modelos de aprendizado de máquina, dados e resultados entre os sistemas.[9]