Deploy de Modelos e MLOps

Meu modelo está com excelentes métricas, e agora?

Apresentação

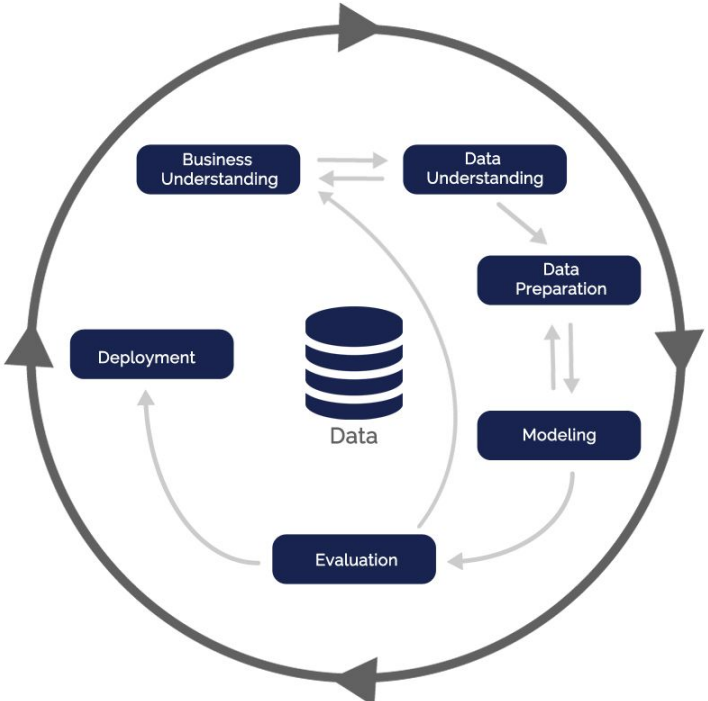
- Ciclo de desenvolvimento em DS
- Formatos de produtização dos Modelos
- MLOps
- Exemplo

Mateus Cichelero da Silva

Engenheiro Eletricista - UFSC

Engenheiro de Machine Learning - Robô Laura

Ciclo de desenvolvimento em DS



The Machine Learning Canvas (v0.4)

Designed for: | Designed by: | Date: | Iteration:

Decisions

How are predictions used to make decisions that provide the proposed value to the end-user?

ML task

Input, output to predict, type of problem.

Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

Data Sources

Which raw data sources can we use (internal and external)?

Collecting Data

How do we get new data to learn from (inputs and outputs)?

Making Predictions

When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?

Offline Evaluation

Methods and metrics to evaluate the system before deployment.

Features

Input representations extracted from raw data sources.

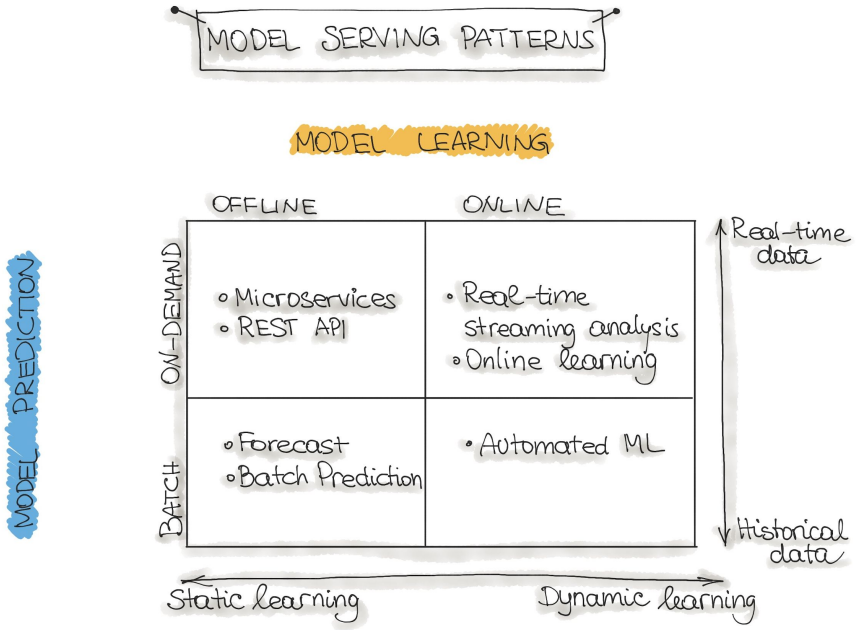
Building Models

When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?

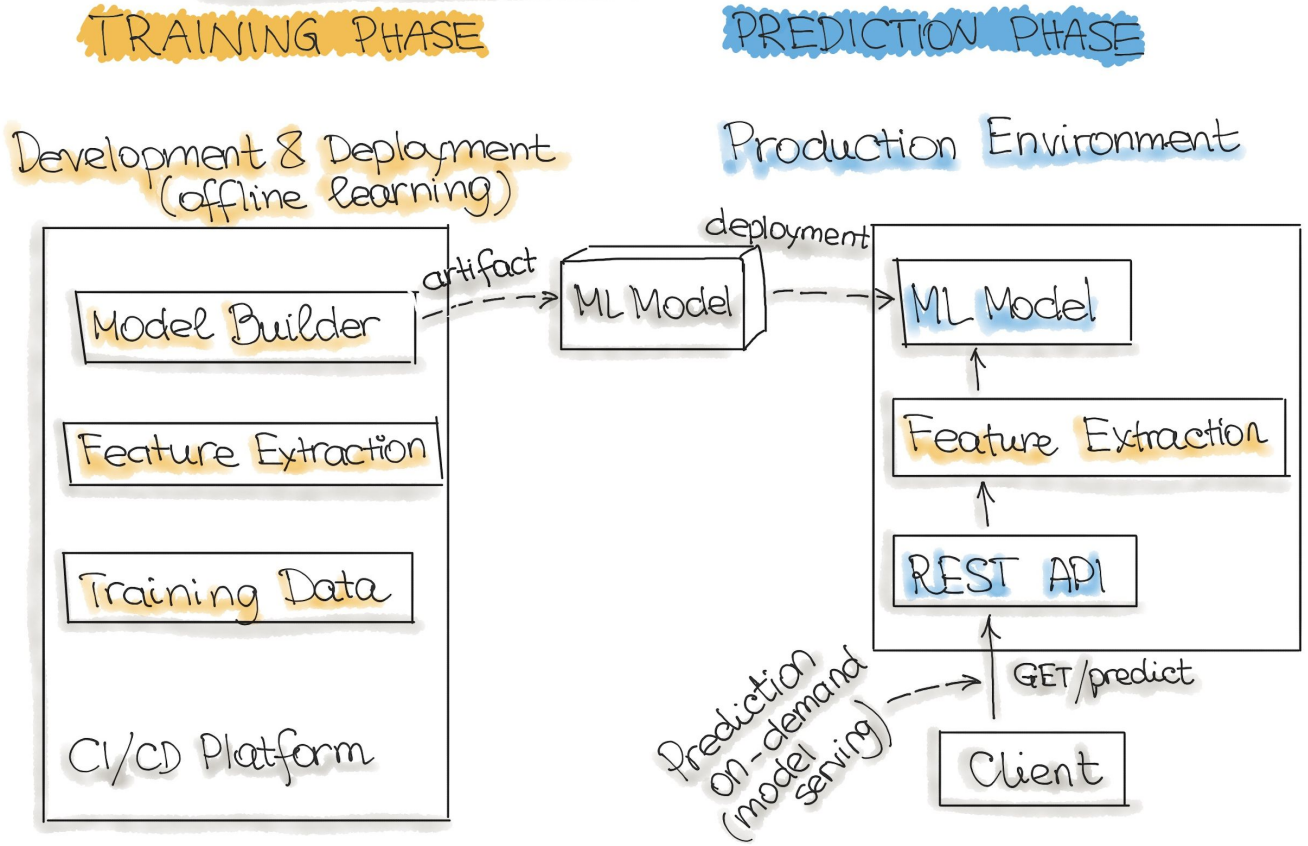
Live Evaluation and Monitoring

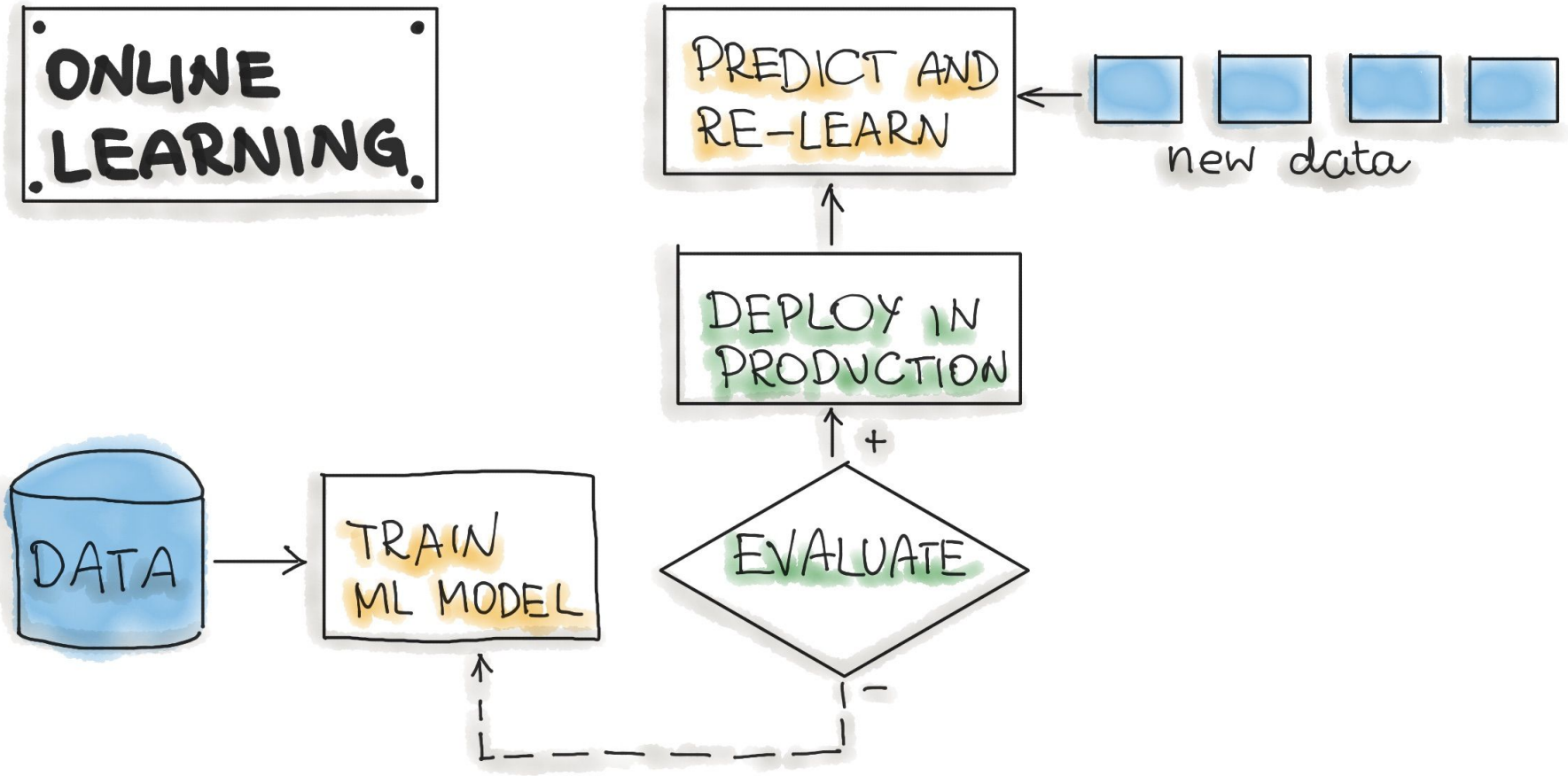
Methods and metrics to evaluate the system after deployment, and to quantify value creation.

Formatos de produtização dos Modelos

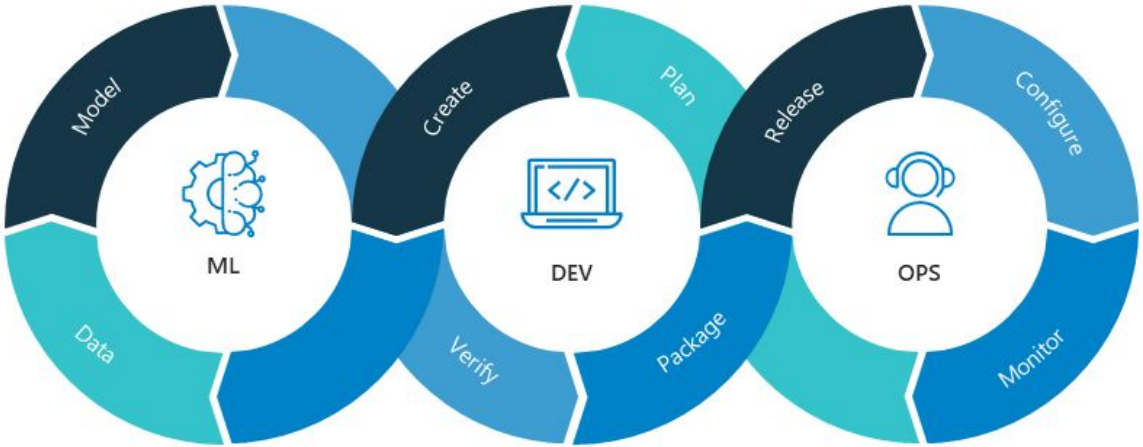


MODEL SERVING AS MICROSERVICE




ONLINE LEARNING
PREDICT AND RE-LEARN
new data
DEPLOY IN PRODUCTION
+
DATA
TRAIN ML MODEL
EVALUATE
-

MLOps


Model
Data
ML
Create
Plan
Verify
Package
DEV
Release
Configure
Monitor
OPS

Exemplo

- Criação de API para disponibilizar resultados de um modelo de classificação - Flask
- Exemplo de API em produção - Robô Laura

	Open-Format	Vendor	File Extension	License	ML Tools & Platforms Support	Human-readable	Compression
"almagination"	–	–	–	–	–	–	✓
PMML	✓	DMG	.pmml	AGPL	R, Python, Spark	✓ (XML)	✘
PFA	✓	DMG	JSON		PFA-enabled runtime	✓ (JSON)	✘
ONNX	✓	SIG LFAI	.onnx		TF, CNTK, Core ML, MXNet, ML.NET	–	✓
TF Serving Format	✓	Google	.pf		Tensor Flow	✘	g-zip
Pickle Format	✓		.pkl		scikit-learn	✘	g-zip
JAR/ POJO	✓		.jar		H2O	✘	✓
HDF	✓		.h5		Keras	✘	✓
MLEAP	✓		.jar/ .zip		Spark, TF, scikit-learn	✘	g-zip
Torch Script	✘		.pt		PyTorch	✘	✓
Apple .mlmodel	✘	Apple	.mlmodel		TensorFlow, scikit-learn, Core ML	–	✓

Muito obrigado! Dúvidas?

mateus.silva@laura-br.com