



Prof. Jean Carlos Alves

DataOps e Implantação de Sistemas de Machine Learning

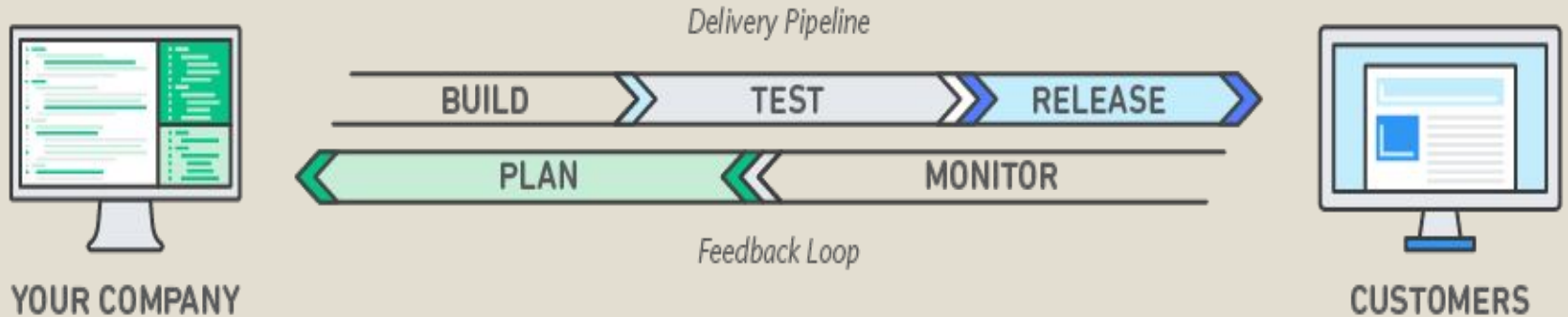


PUC Minas

Devops

- "DevOps é uma metodologia de desenvolvimento de software que traz entrega contínua para o ciclo de vida de desenvolvimento de sistemas, combinando equipes de desenvolvimento e equipes de operações em uma única unidade responsável por um produto ou serviço."

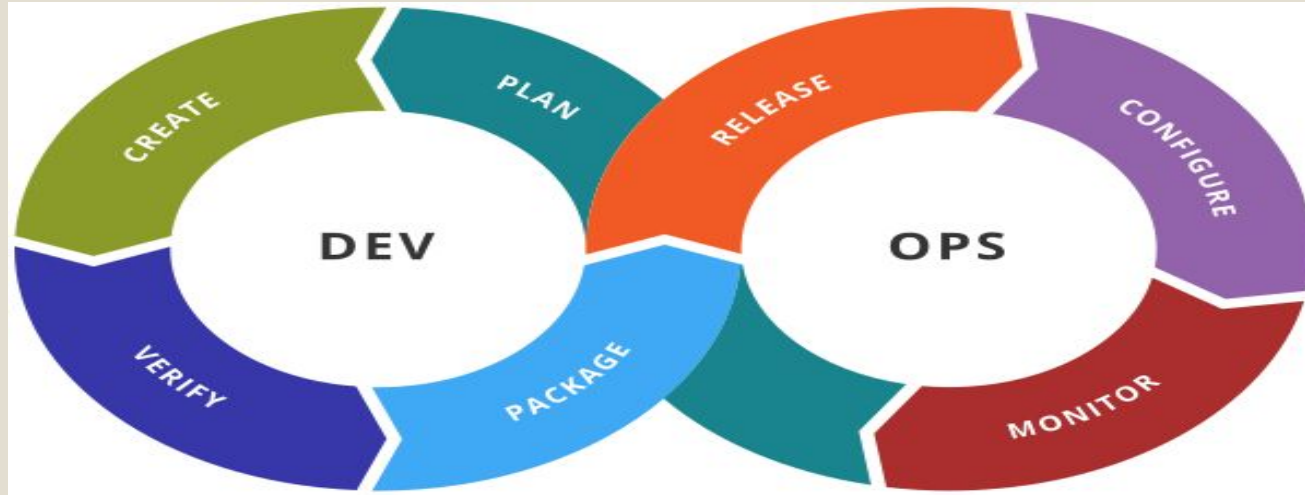
Devops



Benefícios Devops

- Velocidade
- Entrega rápida
- Confiabilidade
- Escala
- Equipes colaborativas
- Segurança

Ciclo de vida Devops



<https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7>

Ciclo de vida Devops

- Código
- Compilação
- Testes automatizados
- Artefatos
- Release
- Configuração
- Monitoramento

Práticas Devops

- Integração Contínua
- Entrega Contínua
- Microsserviços
- Infraestrutura como código
- Monitoramento
- Equipe Colaborativa

DataOps

- "Capacidade de habilitar soluções, desenvolver produtos de dados e ativar dados para valor de negócios em todas as camadas de tecnologia, da infraestrutura à experiência"
Michele Goetz, Vice-Presidente e Analista Principal da Forrester

Princípios DataOps

- Satisfação Cliente
- Análise do trabalho
- Mudanças
- Trabalho em equipe
- Interações diárias
- Auto organizar
- Reduza o Heroísmo

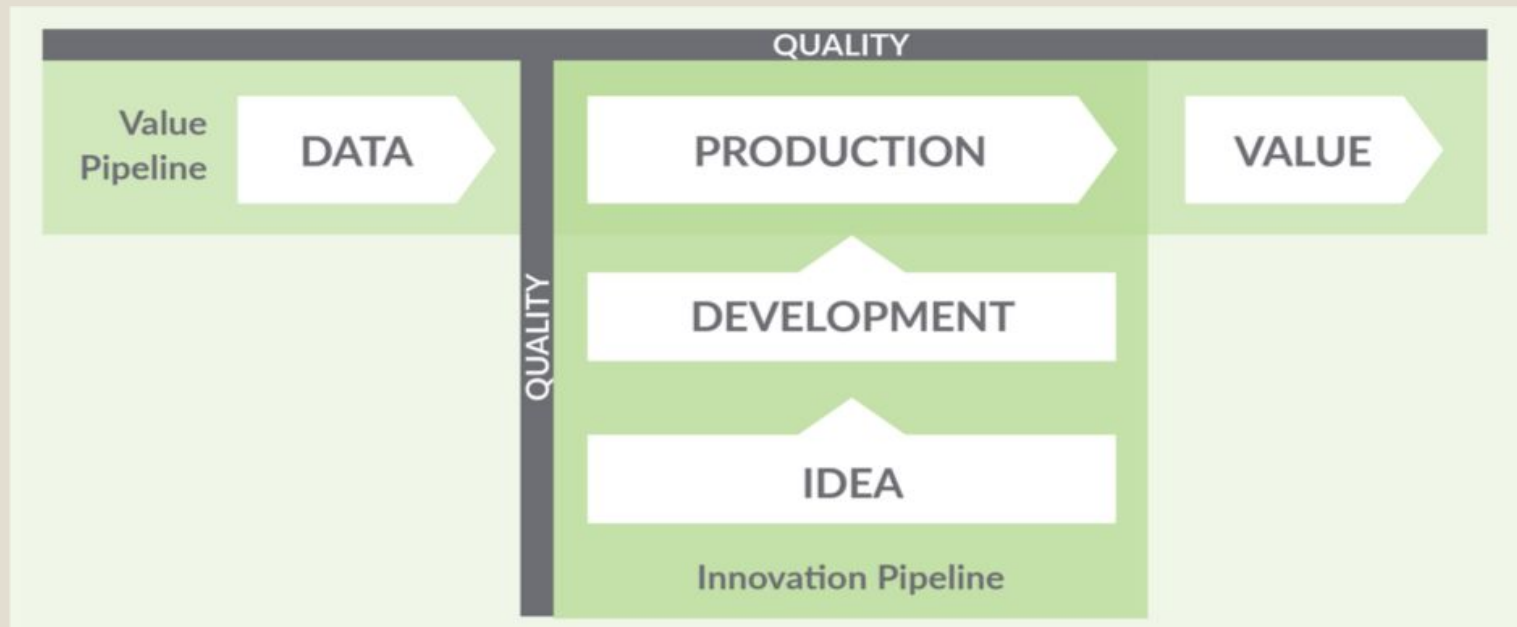
Princípios DataOps

- Reflita
- Analytics é código
- Orquestrar
- Resultados reproduzíveis
- Ambientes descartáveis
- Simplicidade
- Análise é fabricação

Princípios DataOps

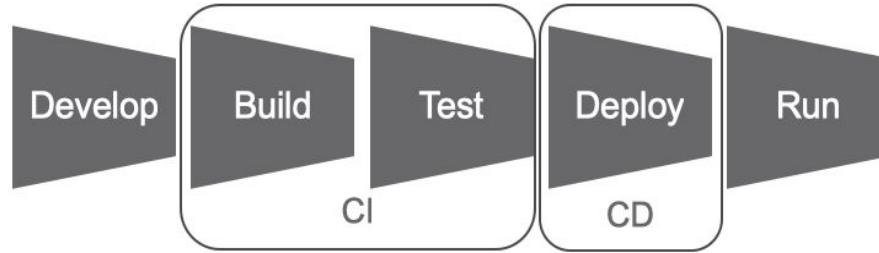
- Qualidade é fundamental
- Monitorar qualidade e desempenho
- Reutilizar componentes
- Otimizar tempo para implantação de soluções

Ciclo de vida DataOps

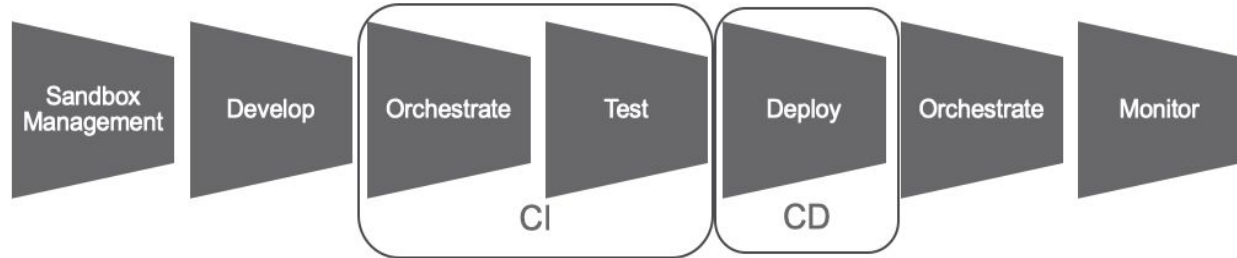


Devops X Dataops

DevOps Process



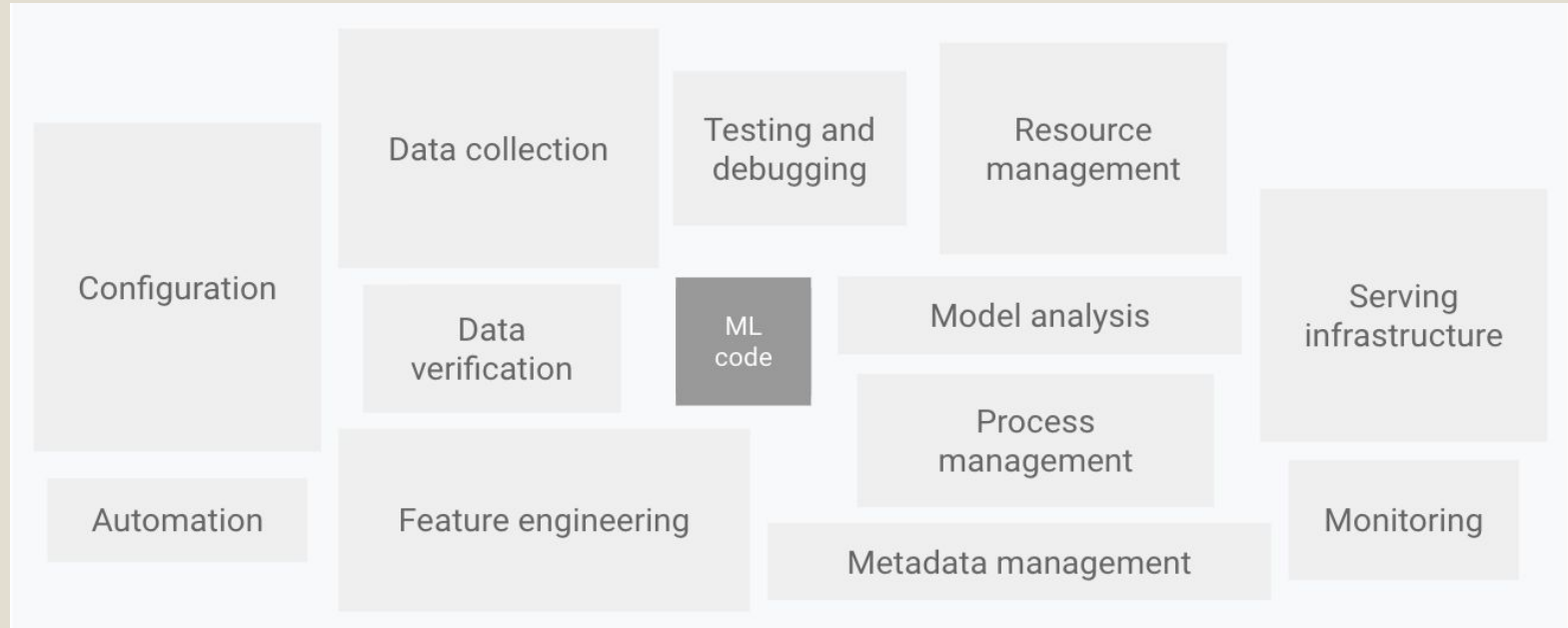
DataOps Process



MLOps

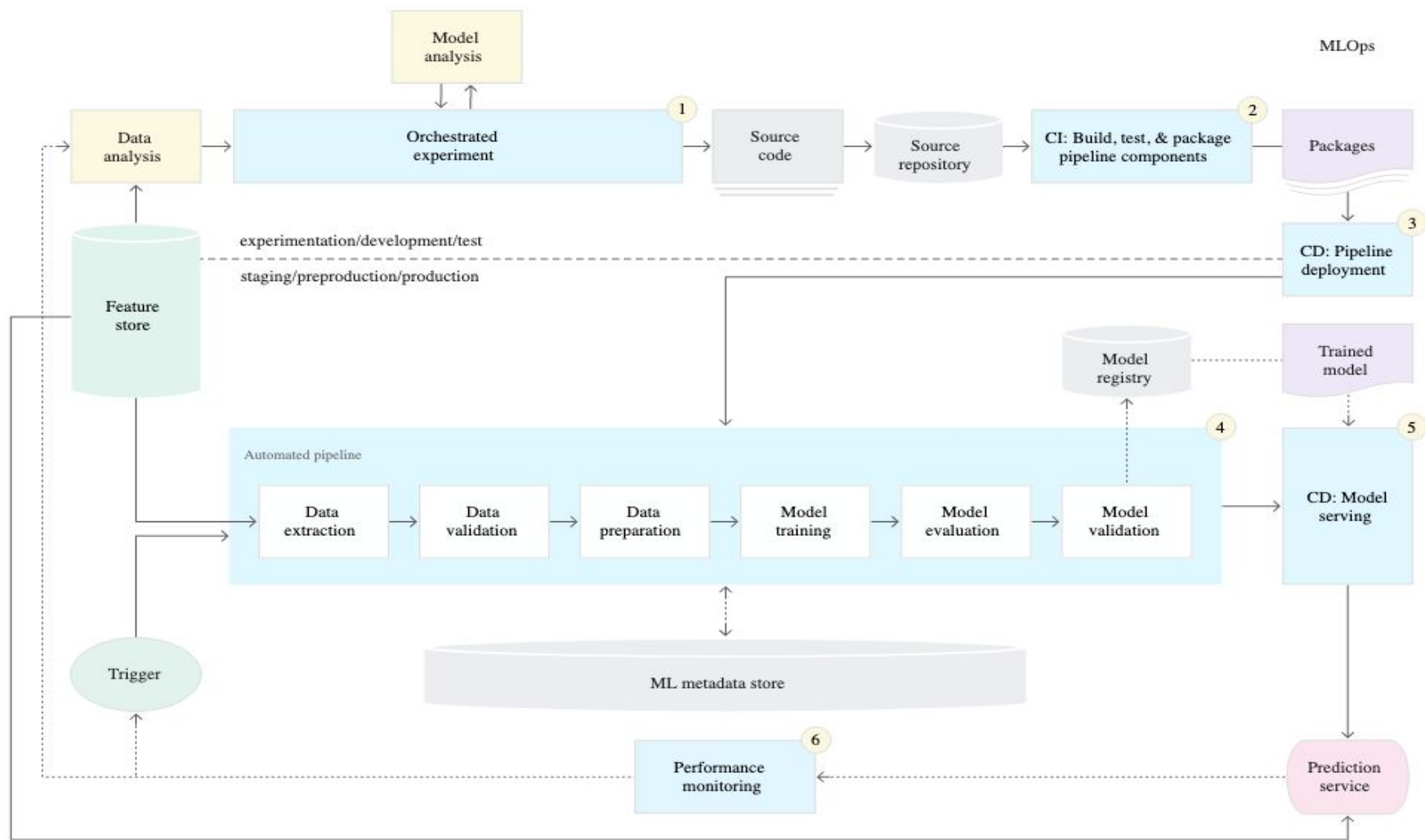
- "MLOps (um composto de aprendizado de máquina e “operações de tecnologia da informação”) é uma nova disciplina / foco / prática para colaboração e comunicação entre cientistas de dados e profissionais de tecnologia da informação (TI) enquanto automatiza e produz algoritmos de aprendizado de máquina. Por meio da prática e das ferramentas, o MLOps visa estabelecer uma cultura e um ambiente onde as tecnologias de ML podem gerar benefícios de negócios ao construir, testar e lançar a tecnologia de ML de forma rápida, frequente e confiável." Nisha Talagala

Elementos sistemas ML



<https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=pt-br>

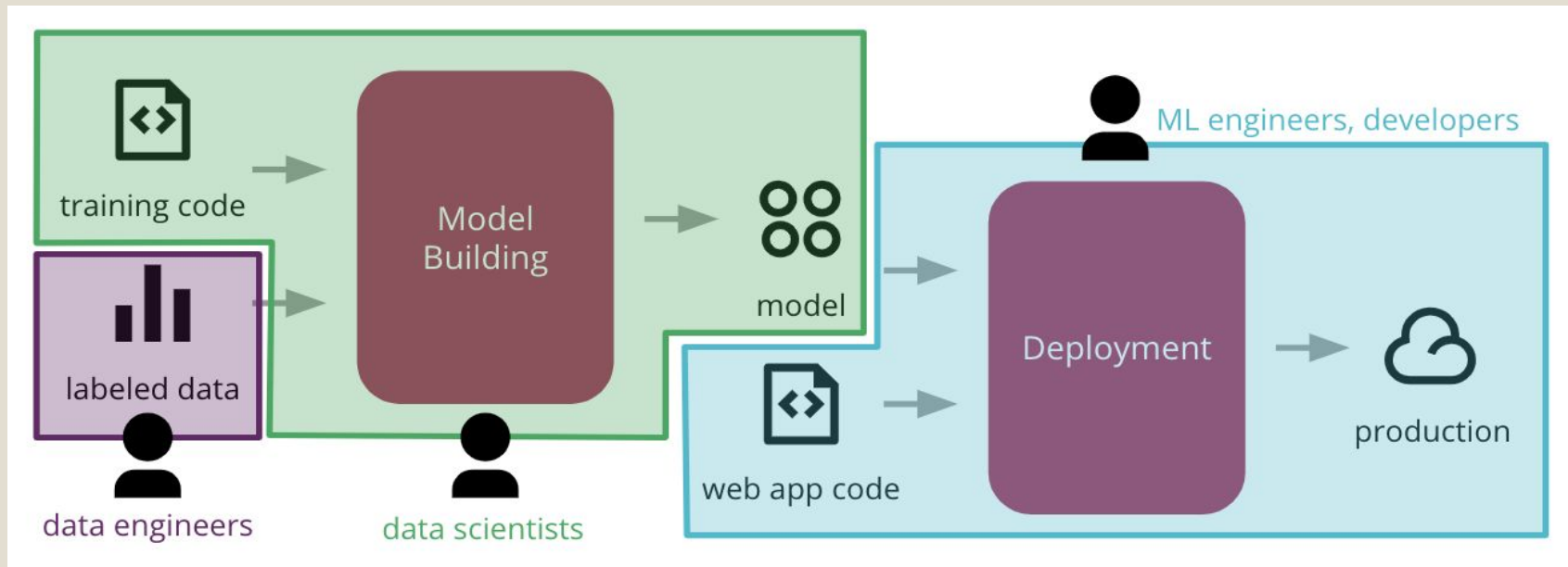
Exemplo pipeline ML



CD4ML - Continuous Delivery for Machine Learning

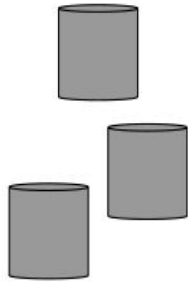
- "A entrega contínua é a capacidade de colocar mudanças de todos os tipos - incluindo novos recursos, mudanças de configuração, correções de bugs e experimentos - em produção ou nas mãos dos usuários com segurança e rapidez de forma sustentável." Jez Humble e Dave Farley

Etapas para criação de modelo ML

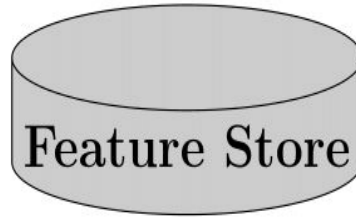


Dados (Feature Store)

Raw/Structured Data



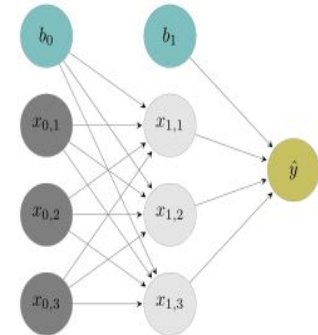
Feature Engineering



Training



Models



Dados (Feature Store)

- Reutilização
- Flexibilidade
- Disponibilidade
- Qualidade

Desenvolvimento (MLFLOW)

- Rastreamento
- Projetos
- Modelos
- Registros de modelos

Desenvolvimento (MLFLOW)

The screenshot displays the MLflow web interface for a specific run. At the top, the MLflow logo is on the left, and 'Github' and 'Docs' links are on the right. The main heading is 'Run 7c1a0d5c42844dcdb8f5191146925174'. Below this, metadata is shown: Experiment Name: Default, Start Time: 2018-06-04 23:47:22, Source: train.py, Git Commit: 3aa48cffe58b8d9d69f5, User: mlflow, and Duration: 145ms.

Two expandable sections are visible: 'Parameters' and 'Metrics'. The 'Parameters' section shows a table with two entries: 'alpha' with value '0' and 'l1_ratio' with value '0'. The 'Metrics' section shows a table with three entries: 'mse' with value '0.578', 'r2' with value '0.288', and 'rmse' with value '0.742'.

Below these is the 'Artifacts' section, which is expanded to show a file tree. The tree includes a 'model' directory containing 'MLmodel' and 'model.pkl'. To the right of the file tree, the full path and size of the selected artifact are shown: 'Full Path: /Users/mlflow/mlflow-prototype/miruns/0/7c1a0d5c42844dcdb8f5191146925174/artifacts/model/MLmodel' and 'Size: 259B'. A detailed JSON-like description of the artifact follows, including its path, flavors (python, data, loader), and sklearn-specific details like the pickled model path, sklearn version (0.19.1), run ID, and creation timestamp.

```
artifact_path: model
flavors:
  python_function:
    data: model.pkl
    loader_module: mlflow.sklearn
  sklearn:
    pickled_model: model.pkl
    sklearn_version: 0.19.1
run_id: 7c1a0d5c42844dcdb8f5191146925174
utc_time_created: '2018-06-05 06:47:22.757025'
```

<https://mlflow.org/docs/latest/tutorials-and-examples/tutorial.html>

Desenvolvimento (SageMaker Stúdio)

The screenshot shows the Amazon SageMaker Studio interface. On the left, the 'Components and registries' sidebar is visible, showing a table of experiments. The main area displays a Jupyter Notebook with Python code for creating an S3 bucket and uploading data files. The code includes comments in Spanish and a confirmation message.

Name	Created	Last modified
framework-mode-trial-202...	5 months ago	5 months ago
framework-mode-trial-202...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago
algorithm-mode-trial-2020...	5 months ago	5 months ago

```

bucket = 'sagemaker-studio-()-().format(sess.region_name, account_id)
prefix = 'xgboost-churn'

try:
    if sess.region_name == "us-east-1":
        sess.client('s3').create_bucket(Bucket=bucket)
    else:
        sess.client('s3').create_bucket(Bucket=bucket,
                                         CreateBucketConfiguration={'LocationConstraint': sess.region_name})

except Exception as e:
    print("Looks like you already have a bucket of this name. That's good. Uploading the data files...")

# Return the URLs of the uploaded file, so they can be reviewed or used elsewhere
s3url = S3Uploader.upload('data/train.csv', 's3://()/()/().format(bucket, prefix, 'train'))
print(s3url)
s3url = S3Uploader.upload('data/validation.csv', 's3://()/()/().format(bucket, prefix, 'validation'))
print(s3url)

Looks like you already have a bucket of this name. That's good. Uploading the data files...
s3://sagemaker-studio-us-east-2-943286545934/xgboost-churn/train/train.csv
s3://sagemaker-studio-us-east-2-943286545934/xgboost-churn/validation/validation.csv
  
```


Desenvolvimento (SageMaker Studio)

random_cuf_forest.ipynb

conda_python3

Computing Anomaly Scores

Now, let's compute and plot the anomaly scores from the entire taxi dataset.

```
[ ]: results = rcf_inference.predict(taxi_data_numpy)
scores = [datum['score'] for datum in results['scores']]

# add scores to taxi data frame and print first few values
taxi_data[scores] = pd.Series(scores, index=taxi_data.index)
taxi_data.head()
```

```
[ ]: fig, ax1 = plt.subplots()
ax2 = ax1.twinx()

#
# »Try this out - change 'start' and 'end' to zoom in on the
# anomaly found earlier in this notebook
#
start, end = 0, len(taxi_data)
#start, end = 5500, 6500
taxi_data_subset = taxi_data[start:end]

ax1.plot(taxi_data_subset['value'], color='C0', alpha=0.8)
ax2.plot(taxi_data_subset['score'], color='C1')

ax1.grid(which='major', axis='both')
ax1.set_ylabel('Taxi Ride (miles)', color='C0')
ax2.set_ylabel('Anomaly Score', color='C1')

ax1.tick_params('y', colors='C0')
ax2.tick_params('y', colors='C1')

ax1.set_ylim(0, 40000)
ax2.set_ylim(min(scores), 1.4*max(scores))
fig.set_figwidth(10)
```

Note that the anomaly score spikes where our eyeball-norm method suggests there is an anomalous data point as well as in some places where our eyeballs are not as accurate.

Below we print and plot any data points with scores greater than 3 standard deviations (approx 99.9th percentile) from the mean score.

```
[ ]: score_mean = taxi_data[scores].mean()
score_std = taxi_data[scores].std()
score_cutoff = score_mean + 3*score_std

anomalies = taxi_data_subset[taxi_data_subset[scores] > score_cutoff]
anomalies
```

The following is a list of known anomalous events which occurred in New York City within this timeframe:

Trial Component Chart

TRIAL COMPONENTS 9 rows selected. Select rows to toggle chart visibility.

Experiment	Trial	Trial Component	Type
Fruits111	Apple111	DEMO-minerva-byo-2019-11-14-04-26-00-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-14-07-13-53-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-14-17-58-13-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-19-18-05-53-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-19-22-10-02-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-19-22-12-54-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-20-17-13-59-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-21-05-21-26-aws-training-job	arn:aws:sagemaker-us-west-2-33...
Fruits111	Apple111	DEMO-minerva-byo-2019-11-21-18-23-16-aws-training-job	arn:aws:sagemaker-us-west-2-33...

CHARTS

test-metric with 1-minute aggregation

Trial Component List

TRIAL COMPONENTS 1 rows selected

Add chart

Actions

Monitor

Status	Experiment	Type	Trial	Trial component	Monitor
Completed	Fruits111	Training job	Apple111	DEMO-minerva-byo-2...	
Completed	Fruits111	Training job	Apple111	DEMO-minerva-byo-2...	
Completed	Fruits111	Training job	Apple111	DEMO-minerva-byo-2...	
Completed	Fruits111	Training job	Apple111	DEMO-minerva-byo-2...	
Completed	Fruits111	Training job	Apple111	DEMO-minerva-byo-2...	

CHART PROPERTIES

Data type

- Time series
- Summary statistics

Chart type

- Bar
- Line
- Scatter plot

X-axis dimension

- Epoch
- Time
- Periods from start

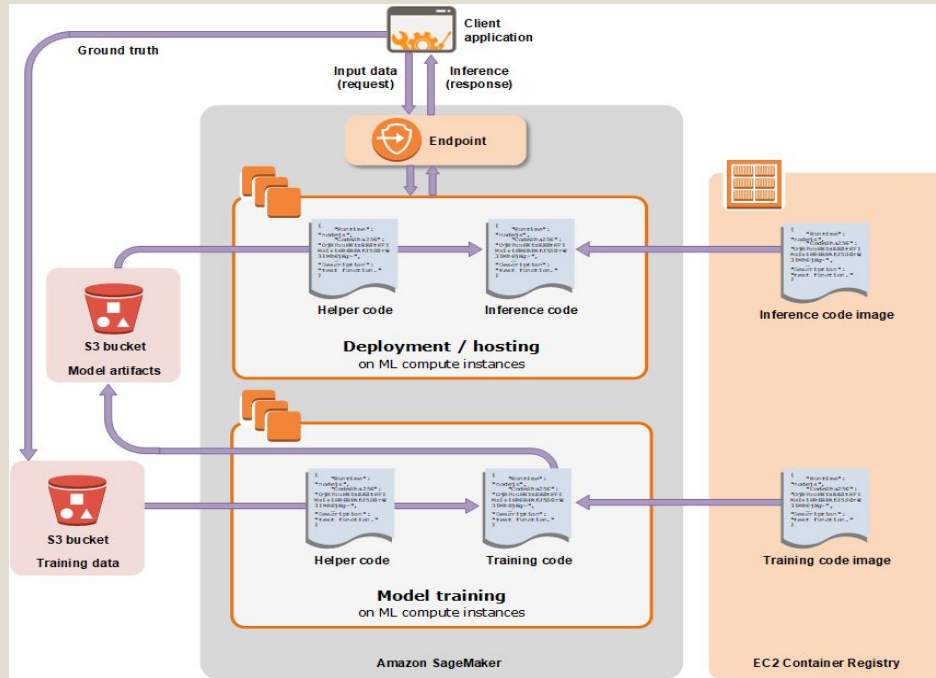
X-axis aggregation

- 1-minute
- 5-minute
- 60-minute

Y-axis

- test-metric: quantitative

Desenvolvimento (SageMaker Stúdio)



<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

Bibliografia

- <https://cio.com.br/tendencias/o-que-e-dataops-analytics-colaborativo-e-multifuncional/>
- <https://aws.amazon.com/pt/devops/what-is-devops/>
- <https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7>
- <https://aws.amazon.com/pt/devops/what-is-devops/>
- <https://www.dataopsmanifesto.org/>
- <https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7>
- <https://www.aitrends.com/machine-learning/mlops-not-just-ml-business-new-competitive-frontier/>
- <https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=pt-br>
- <https://www.aitrends.com/machine-learning/mlops-not-just-ml-business-new-competitive-frontier/>
- <https://martinfowler.com/articles/cd4ml.html>
- <https://hopsworks.readthedocs.io/en/1.1/featurestore/featurestore.html>
- <https://mlflow.org/docs/latest/tutorials-and-examples/tutorial.html>
- <https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>



OBRIGADO



Jean Carlos Alves