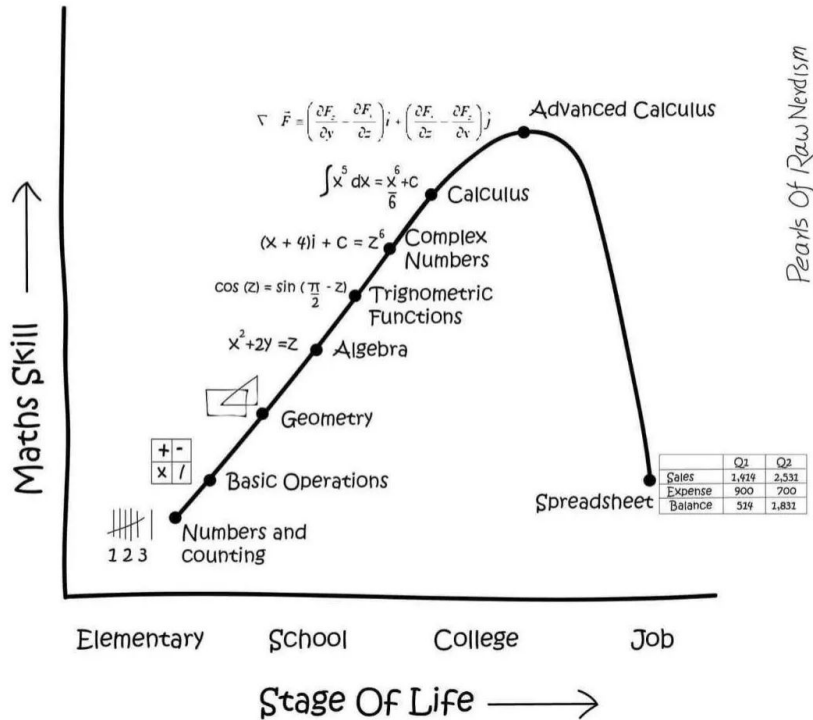


Evitando el desastre utilizando MIFlow

Pydata Madrid Octubre 2023

Ana Sierra & Roger Pou López

La vida del Data Siens



Nos dedicamos a dar vida a modelos

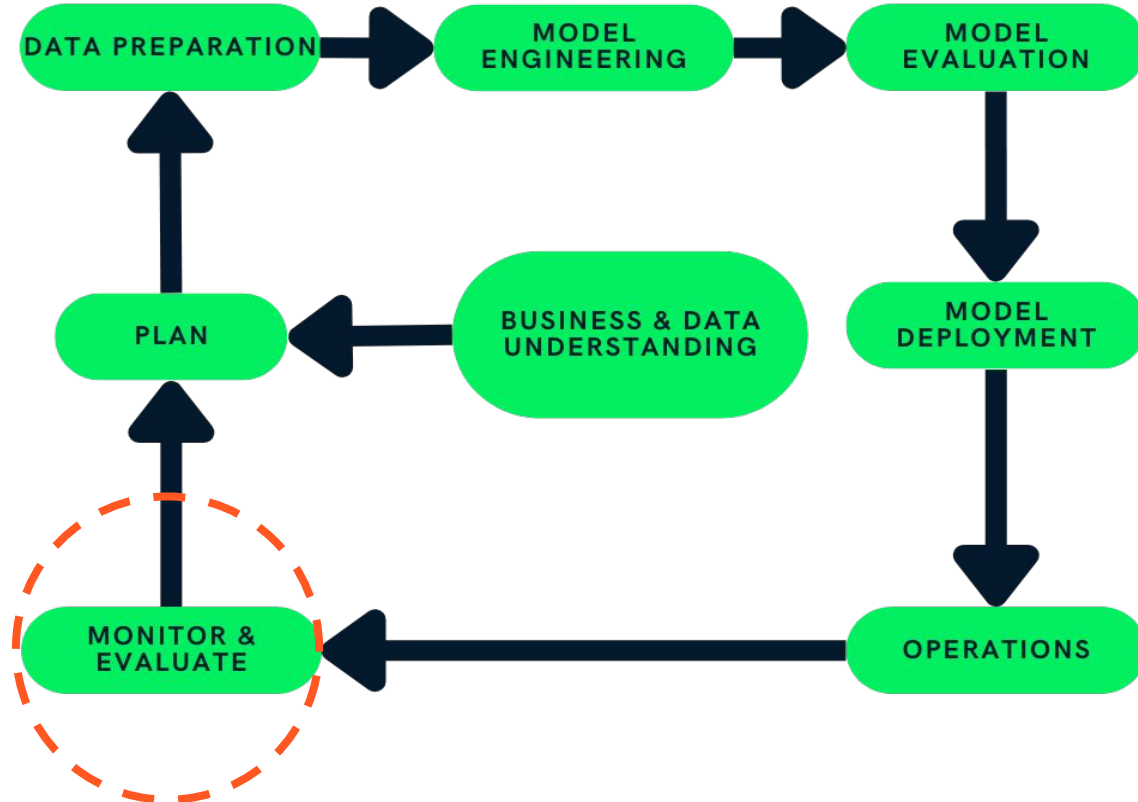
My model on training data



My model on test dataset



Ciclo de vida de los modelos



Si la compañía tiene \$\$\$ y es lo suficientemente grande...

- Contratan DataBricks o alguna plataforma similar con MLOPS
- Contratas a Data Engineers
- Machine Learning Engineers
- Data Scientist (se llevan el mérito)



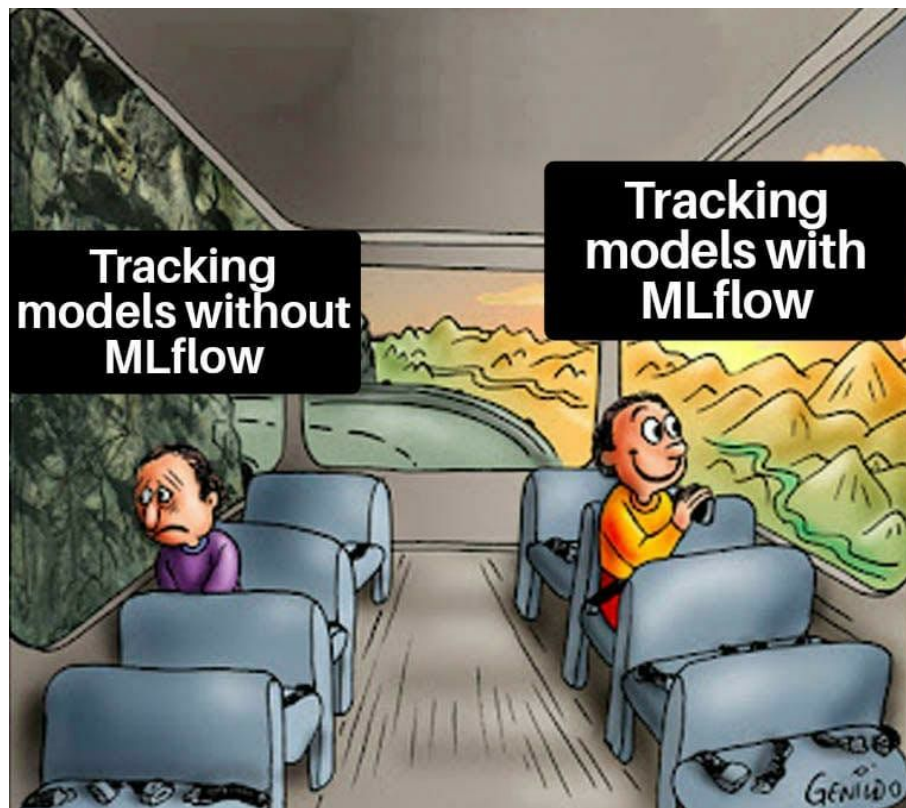
Si es una compañía pequeña y son un poco **ratas**...



mlflow™



Se puede vivir mejor!



¿Qué es Mlflow?

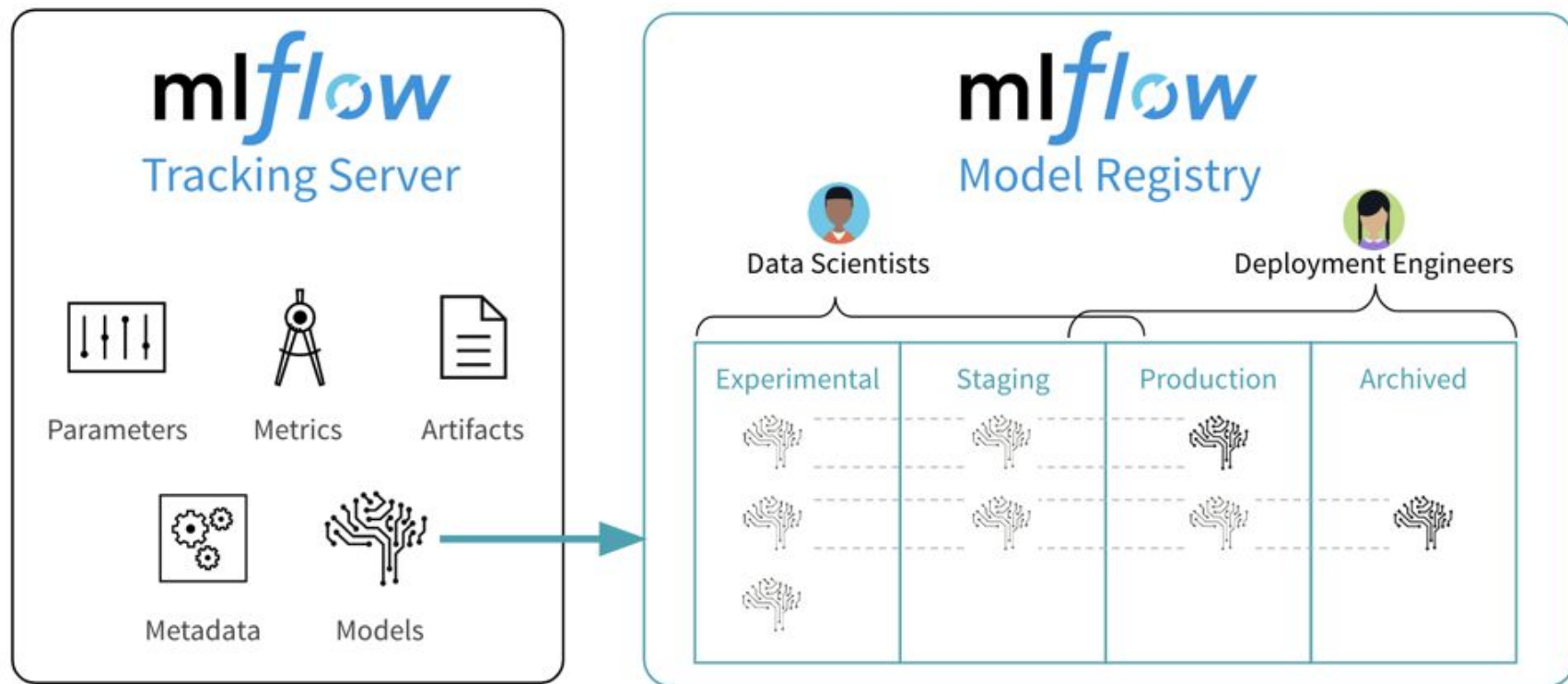
Es una plataforma abierta de aprendizaje automático para gestionar el ciclo de vida de los modelos:

- Funciona con cualquier librería y lenguaje ML
- Se ejecuta de la misma manera en cualquier lugar (cualquier nube...)
- Facilita la reproducibilidad

Mlflow se basa en **tres componentes**:

- **Tracking:** Registro y consulta de experimentos: código, configuraciones, resultados...
- **Proyectos:** Formato de empaquetado para ejecuciones reproducibles en cualquier plataforma
- **Despliegue y/o versionado de Modelos:** Formato general de modelos compatible con diversas herramientas de despliegue

¿Qué es MIFlow?



Tracking con Mlflow

Objetivo: registrar los resultados y parámetros de los modelos para poder compararlos.

Con unas simples líneas de código, se puede realizar un seguimiento de parámetros, métricas y artefactos:

```
import mlflow
# Tracking server
mlflow.set_tracking_uri('http://127.0.0.1:5000')
# Experiment name
mlflow.set_experiment('experiment_name')
# Start run
mlflow.start_run()

# TRAINING
# =====
# For instance:
model.fit(X_train, y_train)
pred = model.predict(X_test)
dictionary_of_metrics = {'mse': mean_squared_error(y_test,
pred),
                        'mae': mean_absolute_error(y_test,
pred)}}
# . . .
```

```
# SAVE MODEL AND OTHER OBJECTS
# =====
# Log model
mlflow.sklearn.log_model(estimator, 'model')
# Log other useful information
mlflow.log_dict(useful_dictionary)
mlflow.log_params(dictionary_of_parameters)
mlflow.log_metrics(dictionary_of_metrics)
mlflow.log_artifacts('path/to/artifacts')
mlflow.evaluate(estimator, dataset, ...)

# End run
mlflow.end_run()
```

Tracking con Mlflow



USER INTERFACE

Displaying Runs from 3 Experiments

Track machine learning training runs in experiments, view runs

Actions: 1. Only show successful 2. Download CSV 3. Select 4. All runs 5. Search 6. Filter 7. Clear

Showing 17 matching runs

	Created	Experiment Name	Duration	Run Name	User	Source	Estimate	Status	Score	Model	F	R	T	Cost (USD)	Run
<input type="checkbox"/>	01-17-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-18-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-19-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-20-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-21-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-22-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-23-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-24-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-25-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-26-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-27-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-28-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-29-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-30-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
<input type="checkbox"/>	01-31-2020	mlflow-experiment	00:00:00	mlflow-experiment	mlflow-experiment	mlflow-experiment	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

API



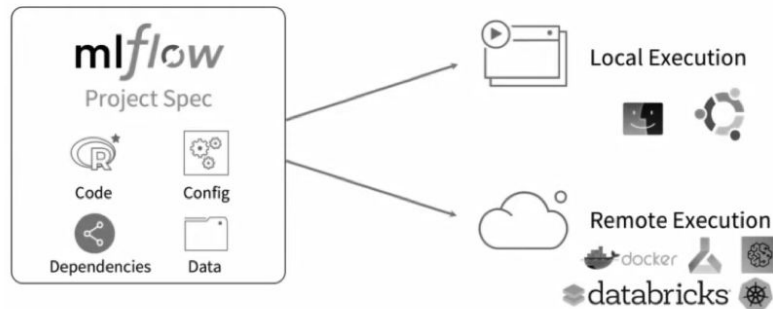
Tracking con Mlflow

mlflow					GitHub	Docs
Comparing 3 Runs						
Run UUID:	d1c0b6387a864aa8873b6ae9fcc215ef 45ac8c41c5db40e394e9b638d02930f0 78962616d1a349cdb432d31c548dfad1					
Start Time:	2018-08-13 15:13:54	2018-08-13 14:34:43	2018-08-13 09:12:03			
Parameters						
Name						
epochs	20	30	20	← Parameters Used for each model experiment		
hidden_layers	3	3	1			
loss_function	mse	binary_crossentropy	binary_crossentropy			
output	32	32	16			
Metrics						
Name						
average_acc	0.878	0.866	0.883	← Metrics Generated for each experiment		
average_loss	0.09	0.441	0.304			
binary_acc	0.977	0.992	0.937			
binary_loss	0.025	0.035	0.212			
validation_acc	0.885	0.879	0.89			
validation_loss	0.025	0.035	0.212			

Proyectos con Mlflow

Objetivo: paquetizar el código de tal forma que sea reproducible.

- Ofrece un formato de empaquetado de código.
- Se integra con el componente de tracking para registrar parámetros.
- Puede integrarse con git para confirmación de código.
- API de ejecución para ejecutar proyectos (Python, R, Java...).
- Soporta ejecución local y remota.



Despliegue y versionado de modelos con Mlflow

Objetivo: Permite gestionar el versionado **de modelos**, así como poner en producción modelos de ML a modo de **endpoint**. Este último, incluye integraciones para hacer el deploy del modelo tanto en Azure ML como en AWS SageMaker.

Registered Models > wine_1

wine_1

Details Serving

Notify me about ☐ Activity on versions I follow

Created Time: 2020-11-13 10:59:00 Last Modified: 2021-08-13 14:42:40

▼ Description [Edit](#)

Wine classifier model
sklearn.ensemble.GradientBoostingClassifier
Team: xyz

► Tags

▼ Versions All Active 2 Compare

<input type="checkbox"/>	Version	Registered at	Created by	Stage	Pending Requests	Description
<input type="checkbox"/>	Version 7	2021-02-08 15:26:21		Staging	1	beta test
<input type="checkbox"/>	Version 6	2021-02-08 14:12:13		Production	—	

Registered Models > wine_1 > Version 7

Version 7

Registered At: 2021-02-08 15:26:21 Creator: Follow Status: Following

Last Modified: 2021-08-13 14:42:40 Source Run: [Run 4bf36e31f9164ef59eb9e761b8d08d3c](#) Stage: Staging

▼ Description [Edit](#)

beta test

▼ Pending Requests

Request	Request by	Actions
Transition to → Production		Approve Reject Cancel

► Tags

▼ Schema

Name	Type
Inputs (12)	
Outputs (1)	

▼ Activities

applied a stage transition None → Staging 1 month ago

¿Cómo podemos usar el Mlflow?

Para usar Mlflow se necesita:

1. **Una máquina virtual / Ordenador.** En esta máquina virtual instalaremos MLflow, podremos ver la UI de MLflow, servirá modelos, etc.
2. **Una base de datos.** La base de datos es el lugar donde MLflow guardará los metadatos del tracking de parámetros. Además, debe ser alguna de las siguientes: MySQL, SQLite o PostgreSQL.
3. **Un lugar donde guardar artefactos:** este es el lugar donde guardaremos los modelos. Se puede usar la propia máquina virtual, aunque lo más típico suele ser usar un Datalake como S3 o Cloud Storage.

Resumen ventajas del uso de MIFlow

- Permite ir guardando las **métricas en una interfaz gráfica**, además de registrarlas en una base datos.
- Permite guardar información sobre las distintas ejecuciones modelos:
 - Feature importance, Betas, curvas ROC, lift...
 - Para el estudio de hyper parámetros de los modelos viene especialmente bien.
 - Permite guardar y comparar parámetros, y características del dataset.
 - Facilita el despliegue de modelos.

Ejemplo de Mlflow

The screenshot shows the Kaggle competition page for 'House Prices - Advanced Regression Techniques'. The header features a banner with the competition title and a 'Join Competition' button. Below the banner, the 'Overview' section includes a description of the competition, a 'Getting Started Notebook' link, and a 'Competition Description' section with an illustration of a row of colorful houses. The right sidebar contains sections for 'Competition Host' (Kaggle), 'Prizes & Awards' (Knowledge), 'Participation' (4,308 Competitors, 4,073 Teams, 18,841 Entries), 'Tags' (Regression, Tabular), and a 'Table of Contents' (Description, Evaluation, Tutorials, Frequently Asked Questions, Citation).

House Prices - Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 4,073 teams · Ongoing

[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#) ...

Overview

CC This competition runs indefinitely with a rolling leaderboard. [Learn more.](#)

Description

Start here if...
You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Getting Started Notebook
To get started quickly, feel free to take advantage of [this starter notebook](#).

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Competition Host
Kaggle

Prizes & Awards
Knowledge
Does not award Points or Medals

Participation
4,308 Competitors
4,073 Teams
18,841 Entries

Tags
[Regression](#) [Tabular](#)

Table of Contents
[Description](#)
[Evaluation](#)
[Tutorials](#)
[Frequently Asked Questions](#)
[Citation](#)

- Es un dataset con un target de regresión
- Vamos a lanzar varios modelos