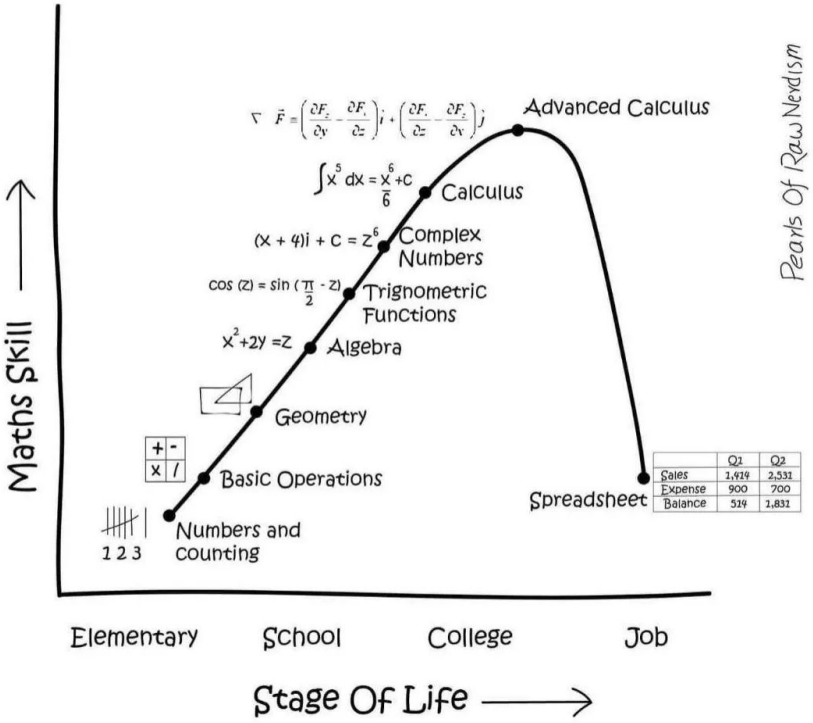
Evitando el desastre utilizando MlFlow

Pydata Madrid Octubre 2023

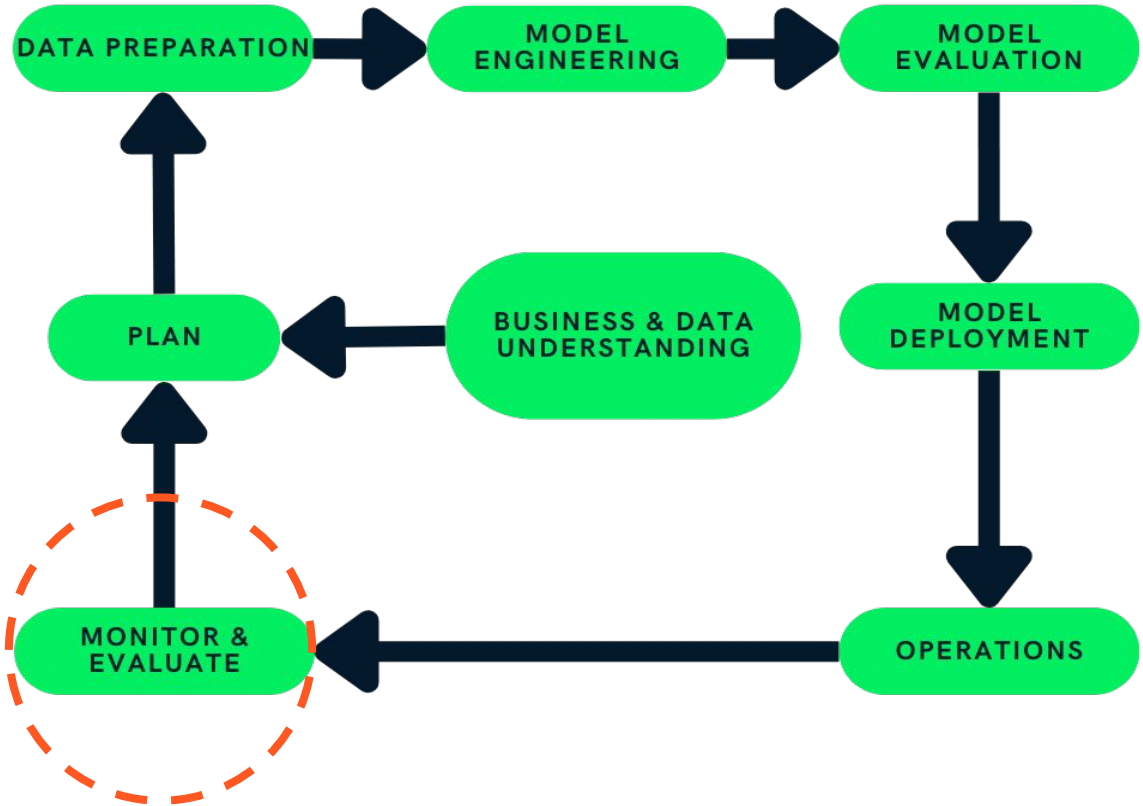
Ana Sierra & Roger Pou López

La vida del Data Siens



Nos dedicamos a dar vida a modelos

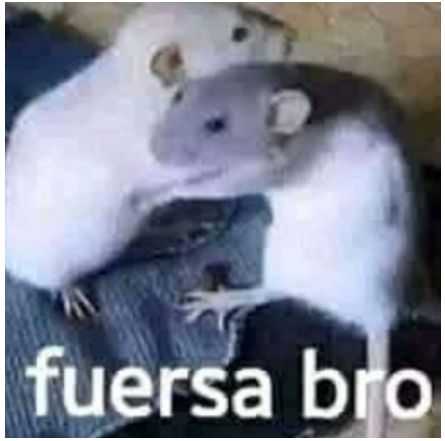
Ciclo de vida de los modelos



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

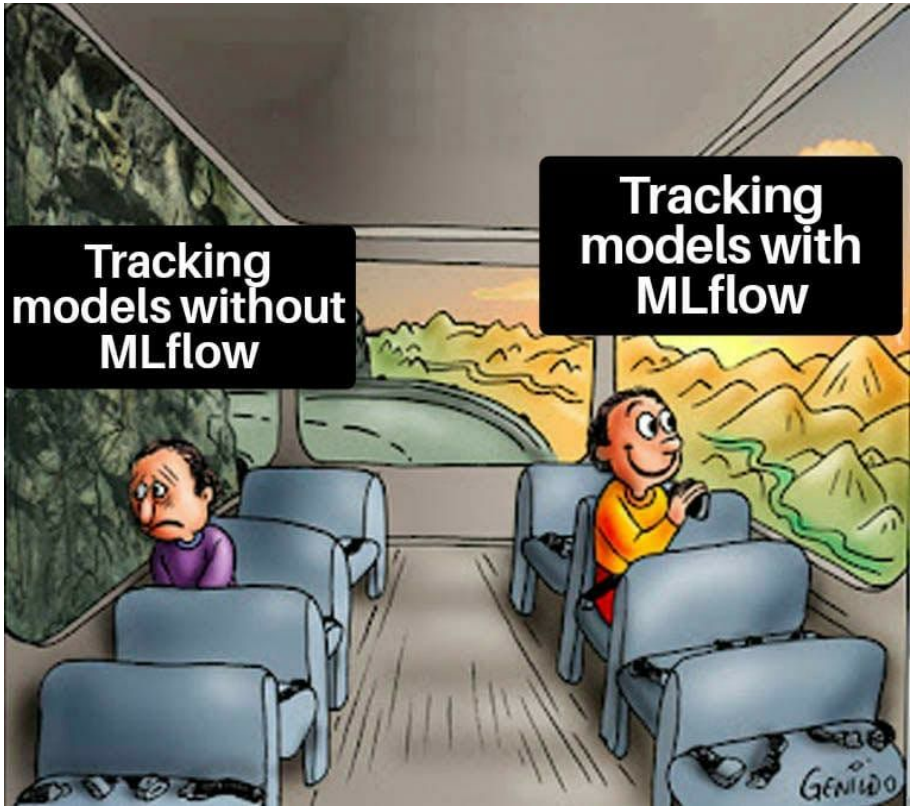
- Contratan DataBricks o alguna plataforma similar con MLOPS
- Contratas a Data Engineers
- Machine Learning Engineers
- Data Scientist (se llevan el mérito)

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



mlflow™

Se puede vivir mejor!



¿Qué es Mlflow?

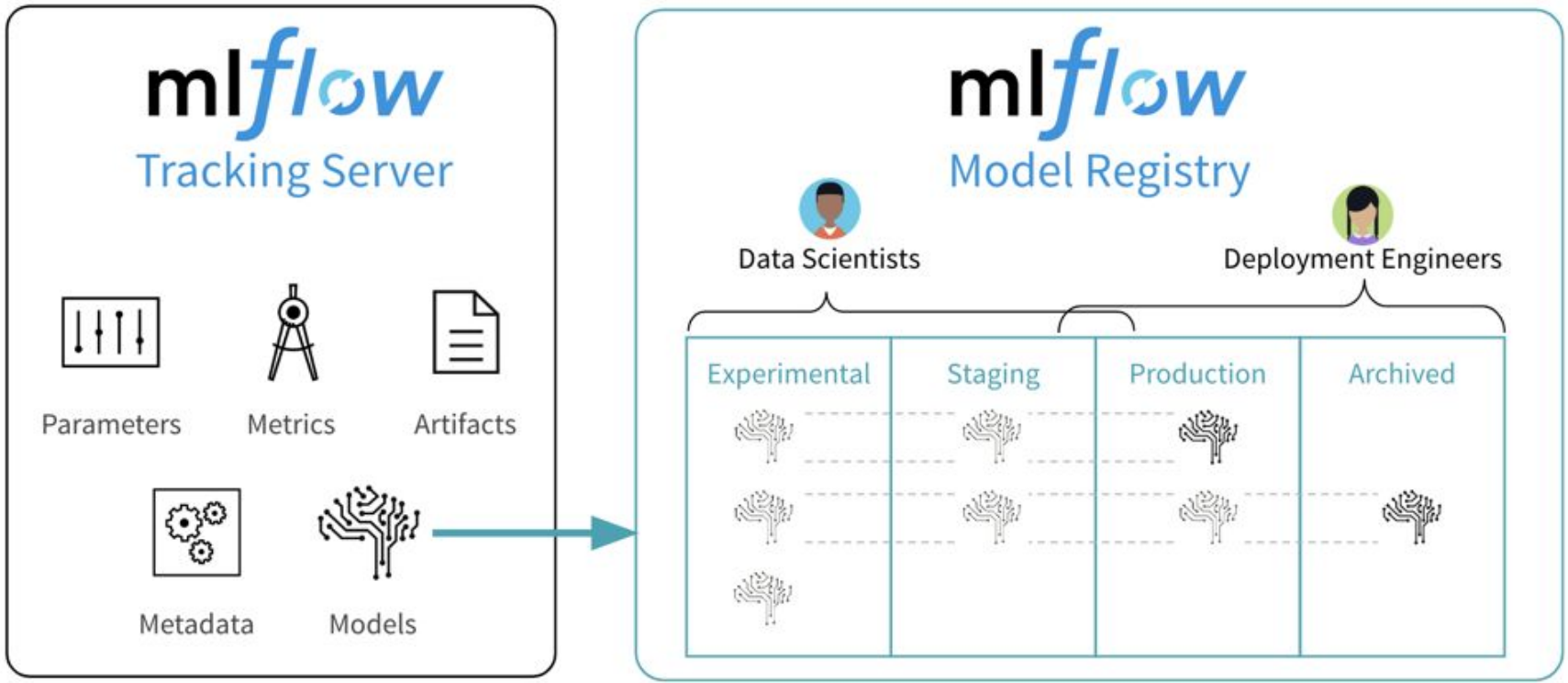
Es una plataforma abierta de aprendizaje automático para gestionar el ciclo de vida de los modelos:

- 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 MlFlow?



Tracking con Mlflow

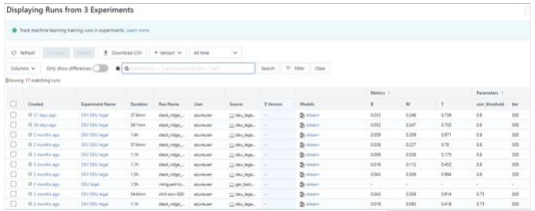
Objetivo: registrar los resultados y parámetros de los modelos para poder compararlos.
Con unas simples líneas de código, se puede realizar un seguimiento de parámetros, métricas y artefactos:

```
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()
```

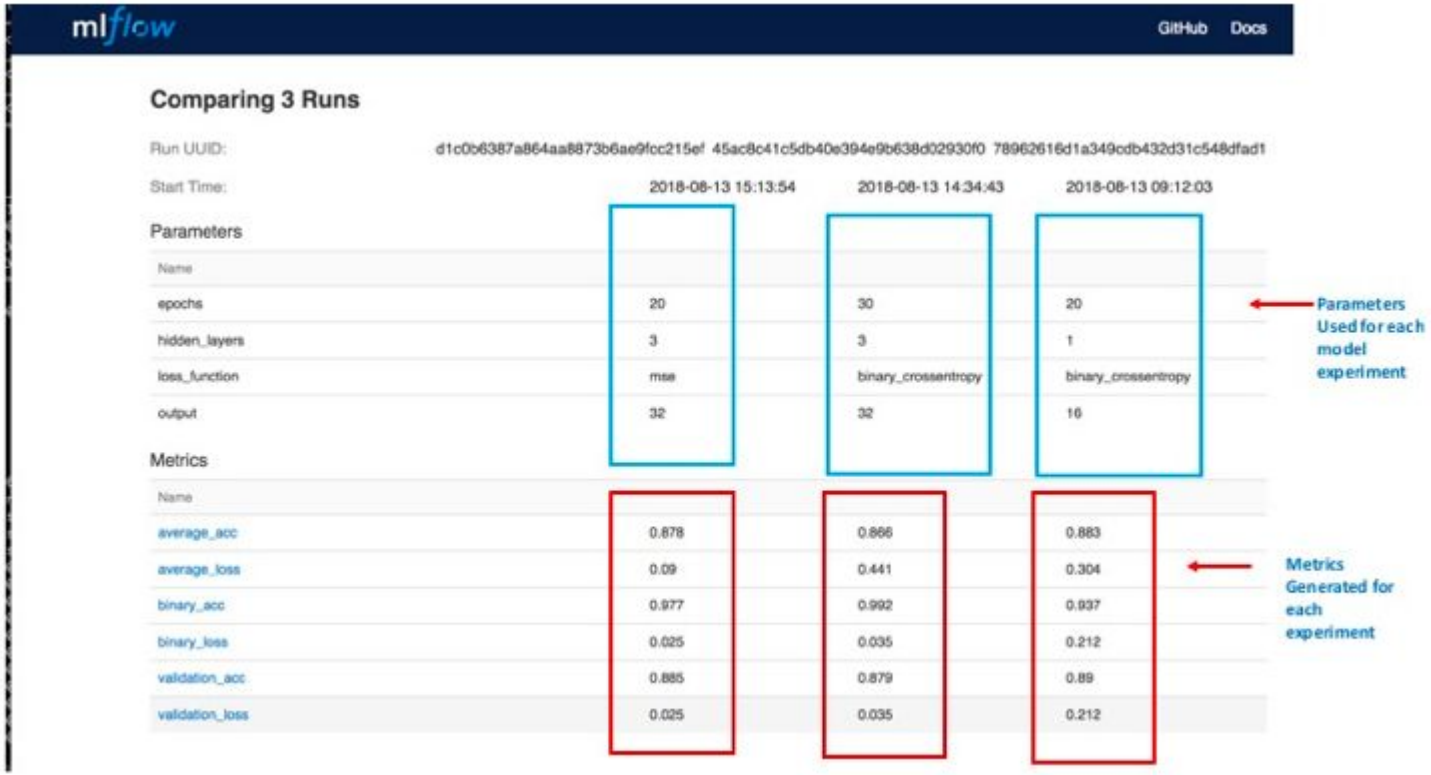
Tracking con Mlflow

USER INTERFACE



API

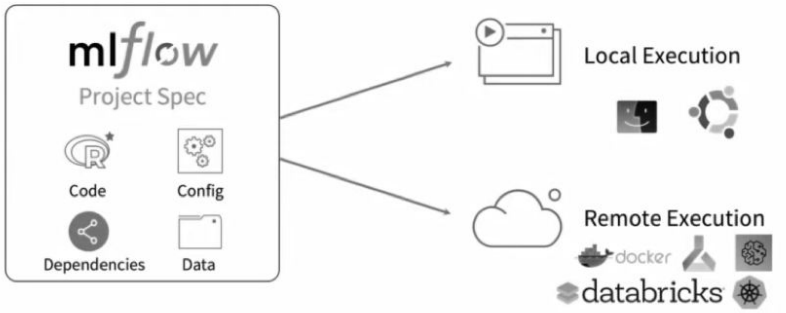
Tracking con Mlflow



Proyectos con Mlflow

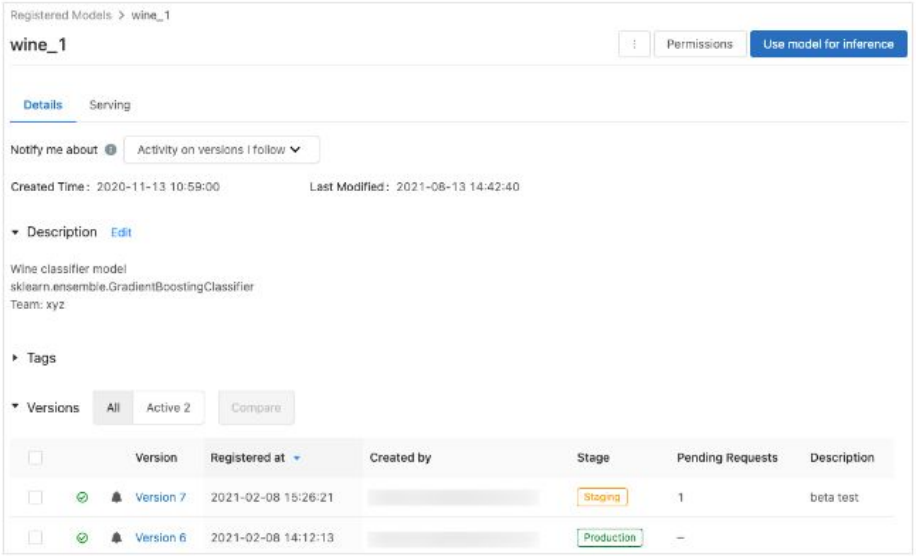
Objetivo: paquetizar el código de tal forma que sea reproducible.

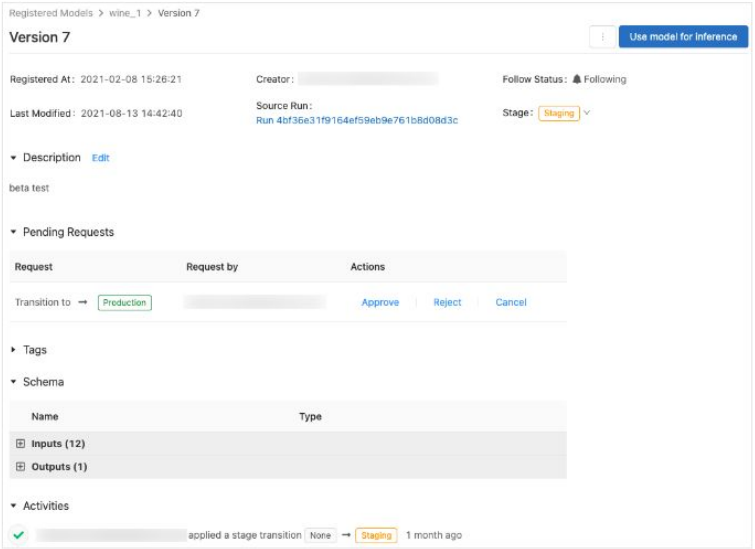
- Ofrece un formato de empaquetado de código.
- Se integra con el componente de tracking para registrar parámetros.
- Puede integrarse con git para confirmación de código.
- API de ejecución para ejecutar proyectos (Python, R, Java...).
- Soporta ejecución local y remota.



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.





¿Cómo podemos usar el Mlflow?

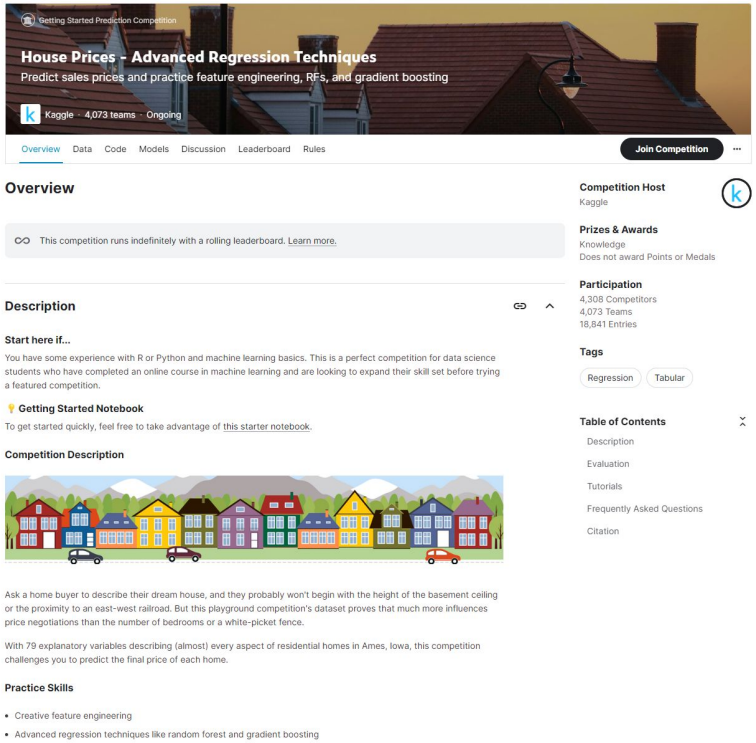
Para usar Mlflow se necesita:

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 MlFlow

- 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



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