

Machine Learning Operations

Chuks Okoli

Last Updated: June 6, 2024

1	Introduction	3
1.1	What is MLOps?	4
1.2	Environment Preparation	6
1.2.1	Configuring Environment with GitHub Codespaces	6
1.2.2	Configuring Environment with Amazon Web Service (AWS)	7
1.3	Model Development	8
1.3.1	Taxi trip prediction with machine learning	8
1.4	MLOps Maturity Model	9
2	Experiment Tracking and Model Management	10
2.1	Introduction to Experiment Tracking	10
2.1.1	How MLflow Helps Manage Machine Learning Experiments	11
2.2	Experiment tracking with MLflow	13
2.3	Machine Learning Lifecycle	14
2.3.1	Logging models in MLflow	14
2.3.2	Model Management	15
2.3.3	Model Registry in Machine Learning	17
2.3.4	Interacting Programmatically with MLflowClient	19
2.4	MLflow in Practice	20
2.4.1	Different scenarios for running MLflow	20
2.4.2	Things to Consider before Configuring MLflow	21
2.5	MLflow: Benefits, limitations and alternatives	21
2.5.1	Remote tracking server	21
2.5.2	Issues with running a remote (shared) MLflow server	21
2.5.3	MLflow limitations (and when not to use it)	22
3	Orchestration and ML Pipelines	23
3.1	Introduction: ML pipelines and Mage	23
3.1.1	Operationalizing ML models	23
3.1.2	Why we need to operationalize ML	23
3.1.3	How Mage helps MLOps	24

3.1.4	Project setup for Mage	24
3.2	Data preparation: ETL and Feature Engineering	25
3.3	Training: sklearn models and XGBoost	25
3.4	Observability: Monitoring and Alerting	25
3.5	Triggering: Inference and Retraining	25
3.6	Deploying: Running operations in Production	25
4	Deployment	26
4.1	Model Deployment	26
4.1.1	Three ways of deploying models	26
4.1.2	Web services: Introduction to Flask	26
4.1.3	Serving the Churn Model with Flask	29
4.1.4	Dependencies and Environment Management: Pipenv	32
4.2	Online Deployment	32
4.2.1	Web services: Deploying models with Flask and Docker	32
4.3	On Cross-Referencing	34
4.4	On Math	34

INTRODUCTION

HELLO THERE, and welcome to Machine Learning Operations. This work is a culmination of hours of effort to create my reference for machine learning operations. All of the explanations are in my own words but majority of the content are based on Alexey Grigorev's DataTalksClub [MLOps Zoomcamp course](#).

EXPLAINING MLOPS TO A 5 YEAR OLD

Imagine you have a magical garden. In this garden, you have a special plant that can grow different kinds of fruits, but you need to take good care of it.

MLOps is like having a magical gardener to help you. This gardener does three important things:

1. **Teaching the Plant:** *The gardener shows the plant pictures of different fruits (like apples, oranges, and bananas) so it can learn to grow them. This is like the gardener helping the plant understand what to do. In MLOps, this is called training a machine learning model.*
2. **Taking Care of the Garden:** *The gardener makes sure the soil is rich, the water is just right, and there are no weeds. The gardener also provides the plant with the best tools and instructions to grow healthy fruits. In MLOps, this is called managing the infrastructure for the machine learning model.*
3. **Sharing the Fruits:** *Once the plant grows the fruits, the gardener picks them and puts them in baskets for everyone to enjoy. The gardener makes sure the fruits are easy to find and delicious. In MLOps, this is called deploying the model so it can be used for something useful.*

Here's the magical part: The gardener watches over the garden to make sure everything is running smoothly. If a bug tries to eat the plant or if the plant stops growing fruits, the gardener fixes the problem right away.

So, MLOps is like having a magical gardener who helps your special plant (machine learning model) stay happy, healthy, and productive, making sure it grows the best fruits for everyone to enjoy!

1.1 What is MLOps?

MLOps, also known as DevOps for machine learning, is an umbrella term that encompasses philosophies, practices, and technologies that are related to implementing machine learning lifecycles in a production environment.

— Microsoft Blog

Machine Learning Operations (MLOps) is a set of best practices for putting machine learning models into production. The process for a machine learning project involves:

- Design - define if machine learning is the right tool for solving the problem
- Train - train the model to find the best possible model
- Operate - deploy the model, and monitor degradation or quality of the model

MLOps is a set of practices for automating everything and working together as a team on a machine learning project.

FUN FACT: MLOps Principles

As machine learning and AI become more common in software, it's important to create guidelines and tools for testing, deploying, managing, and monitoring ML models in real-world use. This is where MLOps comes in. It helps prevent “technical debt” in machine learning projects by ensuring smooth operation and maintenance of models throughout their lifecycle.

MLOps helps to manage and orchestrate the end-to-end machine learning lifecycle by ensuring models are consistently accessible, reproducible, and scalable. It focuses on automating deployment and monitoring of ML pipelines while optimizing the model development process.

The three-phase approach to implementing machine learning (ML) solutions are:

- **Business Understanding and Design:** This phase involves identifying user needs, designing ML solutions to address them, and assessing project develop-

ment. Prioritizing ML use cases and defining data requirements are key steps. The architecture of the ML application, serving strategy, and testing suite are designed based on functional and non-functional requirements.

- **ML Experimentation and Development:** This phase focuses on verifying the applicability of ML by implementing Proof-of-Concept models. It involves iteratively refining ML algorithms, data engineering, and model engineering to deliver a stable, high-quality ML model for production.
- **ML Operations:** Here, the emphasis is on deploying the developed ML model into production using established DevOps practices. Testing, versioning, continuous delivery, and monitoring are essential aspects of this phase.

These phases are interconnected, with design decisions influencing experimentation and deployment options.

Iterative-Incremental Process in MLOps

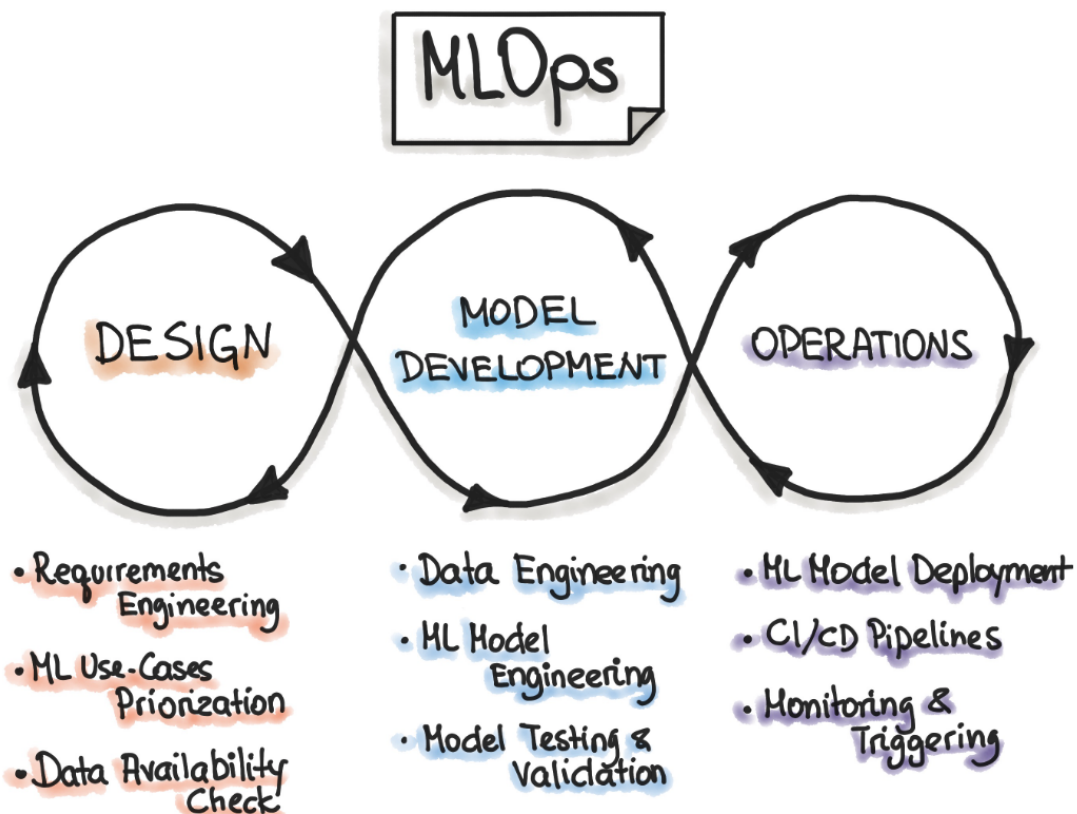


Figure 1.1: The complete MLOps process includes three broad phases of “Designing the ML-powered application”, “ML Experimentation and Development”, and “ML Operations”.

MLOps leverages various tools to simplify the machine learning lifecycle.

- **Machine learning frameworks** like Kubernetes, TensorFlow and PyTorch for model development and training.
- **Version control systems** like Git for code and model version tracking.
- **CI/CD tools** such as Jenkins or GitLab CI/CD for automating model building, testing and deployment.
- **MLOps platforms** like KubeFlow and MLflow manages model lifecycles, deployment and monitoring.
- **Cloud computing platforms** like AWS, Azure and IBM Cloud provide scalable infrastructure for running and managing ML workloads.

1.2 Environment Preparation

1.2.1 Configuring Environment with GitHub Codespaces

To configure the environment using GitHub codespaces, first create a repository on GitHub, give the repository a name, add a “README” file and a `.gitignore` template, choose a license and create the repo. In the repo main page, click on Create codespace on main as shown in **Fig. 1.2**.

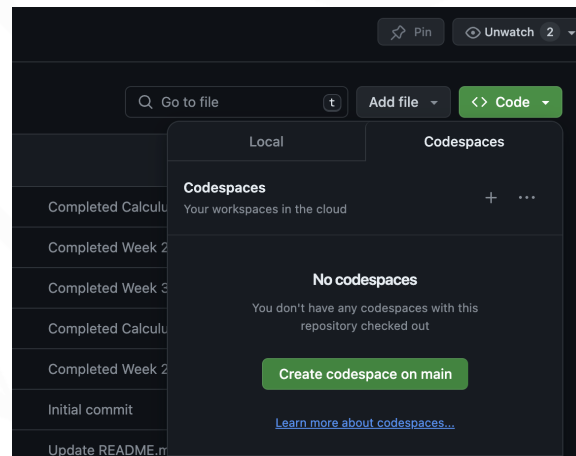


Figure 1.2: GitHub Codespaces setup

A connection to Visual Studio Code can be made through Codespaces. Start a new terminal, change directory to the workspaces and download and install Anaconda distribution of python in it using the step below.

- **Step 1:** Download and install the Anaconda distribution of Python

```
wget https://repo.anaconda.com/archive/Anaconda3-2022.05-Linux-x86_64.sh
```

- **Step 2:** Run this command:

```
bash Anaconda3-2022.05-Linux-x86_64.sh
```

After installing Anaconda, initialize it. In a new terminal, confirm that Anaconda is running in the workspaces

```
(base) @your_username -> /workspaces/ml-ops-zoomcamp-2024 (main) $  
which python  
/home/codespace/anaconda3/bin/python
```

- **Step 3:** Make sure to install pyarrow in order to download parquet file

```
!pip install pyarrow
```

- **Step 4:** Run jupyter notebook

```
jupyter notebook
```

1.2.2 Configuring Environment with Amazon Web Service (AWS)

To install using AWS, create an account in AWS, go to EC2 and create an instance. Select the OS to use e.g. Ubuntu with 64-bit(x86) architecture. Select the instance type that will be sufficient for your project. Configure the cloud resources and launch. Recommended development environment: Linux

- **Step 1:** Download and install the Anaconda distribution of Python

```
wget https://repo.anaconda.com/archive/Anaconda3-2022.05-Linux-x86_64.sh  
bash Anaconda3-2022.05-Linux-x86_64.sh
```

- **Step 2:** Update existing packages

```
sudo apt update
```

- **Step 3:** Install Docker

```
sudo apt install docker.io
```

To run docker without sudo:

```
sudo groupadd docker  
sudo usermod -aG docker $USER
```

- **Step 4:** Install Docker Compose
Install docker-compose in a separate directory

```
mkdir soft  
cd soft
```

To get the latest release of Docker Compose, go to <https://github.com/docker/compose> and download the release for your OS.

```
wget https://github.com/docker/compose/releases/download/v2.5.0/  
docker-compose-linux-x86_64 -O docker-compose
```

Make it executable

```
chmod +x docker-compose
```

Add the soft directory to PATH. Open the .bashrc file with nano:

```
nano ~/.bashrc
```

In .bashrc, add the following line:

```
export PATH="$HOME/soft:${PATH}"
```

Save it and run the following to make sure the changes are applied:

```
source ~/.bashrc
```

- **Step 5: Run Docker**

```
docker run hello-world
```

1.3 Model Development

1.3.1 Taxi trip prediction with machine learning

EXAMPLE 1.1: TLC Trip Prediction

In this section, we use machine learning to predict the trip duration for trips in New York. The data used were collected from the [NYC Taxi and Limousine Commission \(TLC\)](#). This dataset contains yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data is stored in the PARQUET format. We built a simple linear regression model and save the model as a pickle file for subsequent use later one. See the [duration prediction notebook](#).

We can download parquet file using:

```
# download parquet data  
!wget https://d37ci6vzurychx.cloudfront.net/trip-data/  
green_tripdata_2021-01.parquet
```


1.4 MLOps Maturity Model

The MLOps maturity model helps clarify the Development Operations (DevOps) principles and practices necessary to run a successful MLOps environment. The maturity is the extent to which MLOps is implemented in a team. The MLOps maturity model encompasses five levels of technical capability.

Level	Description	Highlights	Technology
0	No MLOps	<ul style="list-style-type: none"> • Hard to manage lifecycle • Disparate teams • Systems as “black boxes” 	<ul style="list-style-type: none"> • Manual builds/deployments • Manual testing • No centralized tracking
1	DevOps but no MLOps	<ul style="list-style-type: none"> • Less painful releases • Limited feedback • Hard to reproduce results 	<ul style="list-style-type: none"> • Automated builds • Automated tests for application code
2	Automated Training	<ul style="list-style-type: none"> • Managed, traceable training • Easy to reproduce model • Low friction releases 	<ul style="list-style-type: none"> • Automated model training • Centralized tracking of model training performance • Model management
3	Automated Deployment	<ul style="list-style-type: none"> • Releases are low friction and automatic • Full traceability from deployment back to original data • Entire environment managed: train > test > production 	<ul style="list-style-type: none"> • Integrated A/B testing • Automated tests for all code • Centralized tracking
4	Full MLOps Automated Operations	<ul style="list-style-type: none"> • Fully automated, monitored • Systems provide improvement information • Zero-downtime 	<ul style="list-style-type: none"> • Automated training/testing • Centralized metrics

Table 1.1: Machine Learning operations maturity model (culled from [Microsoft Blog](#))

EXPERIMENT TRACKING AND MODEL MANAGEMENT

MLOps is critical for the success of AI projects because it allows teams to iterate quickly and deploy machine learning models reliably and at scale.

— Andrew Ng

2.1 Introduction to Experiment Tracking

FUN FACT: MLflow

MLflow is an open-source platform that helps manage the end-to-end machine learning lifecycle. It provides a set of tools to streamline and automate various stages of the machine learning process, from experimentation to deployment.

Machine Learning experiment is the process of building an ML model. Experiment run represents each trial in an ML experiment. A run artifact is any file associated with an ML run. The experiment metadata is all the information related to the experiment. *Experiment tracking* is the process of keeping track of all the **relevant information** from an **ML experiment**, which includes:

- Source code
- Environment
- Data
- Models
- Hyperparameters
- Metrics

WHY IS EXPERIMENT TRACKING SO IMPORTANT

Experiment tracking is important because of **Reproducibility, Organization, and Optimization**

Experiment tracking is done using MLflow. MLflow is “an open source platform for the machine learning lifecycle”. In practice, it’s just a Python package that can be installed with pip, and it contains four main modules:

- Tracking
- Models
- Model Registry
- Projects

The MLflow Tracking module allows you to organize your experiments into runs, and to keep track of:

- Parameters
- Metrics
- Metadata
- Artifacts
- Models

Along with this information, MLflow automatically logs extra information about the run:

- Source code i.e., the name of the file that was used to run the experiment
- Version of the code (git commit)
- Start and end time
- Author

2.1.1 How MLflow Helps Manage Machine Learning Experiments

Here are some key ways in which MLflow can help with machine learning experiments:

Experiment Tracking

MLflow Tracking allows you to log and query experiments using APIs or a web-based interface. Key features include:

- **Log Parameters and Metrics:** Easily log hyperparameters, metrics, and artifacts (such as model files) for each run.
- **Organize and Compare Runs:** Organize runs by experiments and compare results visually.

- **Search and Query:** Search and filter experiments using a web UI or API, allowing you to quickly find runs with specific attributes.

Reproducibility

MLflow promotes reproducibility of experiments by providing:

- **Code Versioning:** Integrates with version control systems (like Git) to log the version of the code that produced a particular run.
- **Environment Management:** Capture and reproduce the software environment using Conda environments or Docker images.
- **Artifacts:** Store and retrieve artifacts such as datasets, models, and images, ensuring all elements of an experiment are preserved.

Model Management

MLflow Models facilitate model packaging, sharing, and deployment:

- **Standardized Format:** Save models in a standardized format that includes the model itself along with its dependencies and environment.
- **Multi-Platform Support:** Export models to various formats (e.g., TensorFlow, PyTorch, scikit-learn) and deploy them to different environments (e.g., cloud services, Docker).
- **Model Registry:** Register and version models, track model lineage, and transition models through stages (e.g., “staging”, “production”).

Deployment

MLflow provides tools for deploying models to various platforms:

- **Built-in Deployment Options:** Deploy models to cloud platforms like AWS SageMaker, Azure ML, and Google Cloud ML Engine.
- **Custom Deployments:** Create custom deployment logic using the MLflow REST API or the command-line interface (CLI).
- **Batch and Real-Time Serving:** Support for both batch and real-time serving of models, enabling various deployment scenarios.

Collaboration

MLflow facilitates collaboration among team members by providing:

- **Centralized Tracking Server:** A centralized tracking server where team members can log and view experiment runs.

- **Sharing and Collaboration:** Share results, models, and insights easily within the team, fostering better collaboration and knowledge sharing.
- **Integration with CI/CD:** Integrate MLflow with continuous integration and continuous deployment (CI/CD) pipelines for automated testing and deployment of models.

Scalability

MLflow is designed to scale with your needs:

- **Distributed Tracking:** Support for tracking experiments across distributed environments, making it suitable for large-scale machine learning projects.
- **Flexible Storage Options:** Use various backend storage systems for logging data, including file systems, databases, and cloud storage.

By integrating MLflow into your machine learning workflow, you can enhance the management, reproducibility, and deployment of your experiments. It provides a comprehensive set of tools that streamline the entire lifecycle of machine learning models, making it easier to track, reproduce, and deploy models at scale.

2.2 Experiment tracking with MLflow

To start experiment tracking, we need to create a conda environment for tracking experiment and activate it.

```
conda create -n exp-tracking-env python=3.9
conda activate exp-tracking-env
```

Once the environment is activated, install the “requirements.txt” file using:

```
pip install -r requirements.txt
```

Check the installed packages using:

```
pip list
```

To launch mlflow, and store all the artifacts in an sqlite database, we run:

```
mlflow ui --backend-store-uri sqlite:///mlflow.db --port 5001
```

If you encounter error when launching mlflow, use:

```
ps -A | grep unicorn
```

to view the processes using the port and kill them with:

```
kill <process number>
```

before re-launching mlflow.

To access mlflow ui, open [in](#) your browser. The mlflow interface is seen in **Fig. 2.1**.

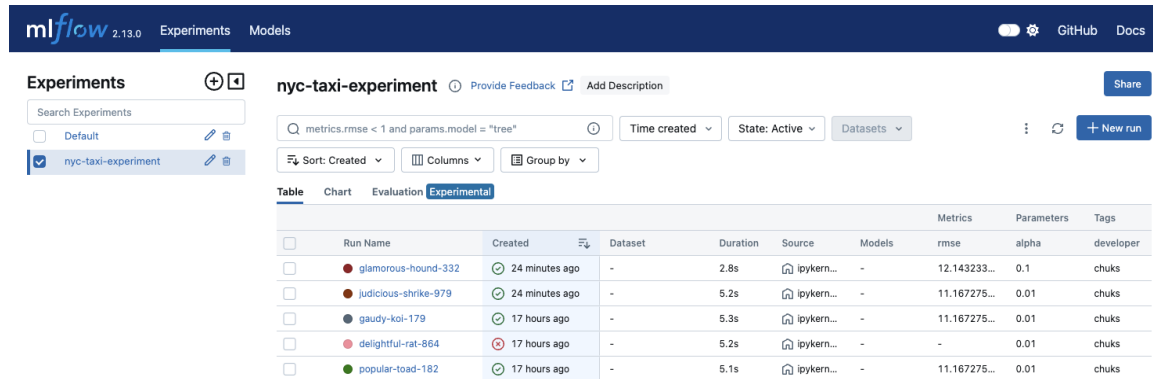


Figure 2.1: MLflow Experiment Interface

The updated notebook with experiment tracking using MLflow and logged predictions in MLflow UI is shown in this [notebook for experiment tracking with mlflow](#).

2.3 Machine Learning Lifecycle

FUN FACT: MLOps cycle

The Machine Learning lifecycle refers to the multiple steps that are needed to build and maintain a machine learning model.

In Machine Learning Lifecycle, we train some model, we tuned the hyperparameters, we evaluated the model and then logged some metrics, hyperparameters and other information needed to mlflow. Once we finish with this experiment tracking stage, it means that we are happy with the model. Then we can start thinking of saving this model and have some kind of versioning. After that, we would like to deploy the model. We may realize that the model needs to be updated in order to scale. Finally, once we deploy the model, the prediction monitoring stage starts.

2.3.1 Logging models in MLflow

Two options :

- Log model as an artifact

```
mlflow.log_artifact("<mymodel>", artifact_path="models/")
```

- Log model using the method "log_model"

```
mlflow.<framework>.log_model(model, artifact_path="models/")
```

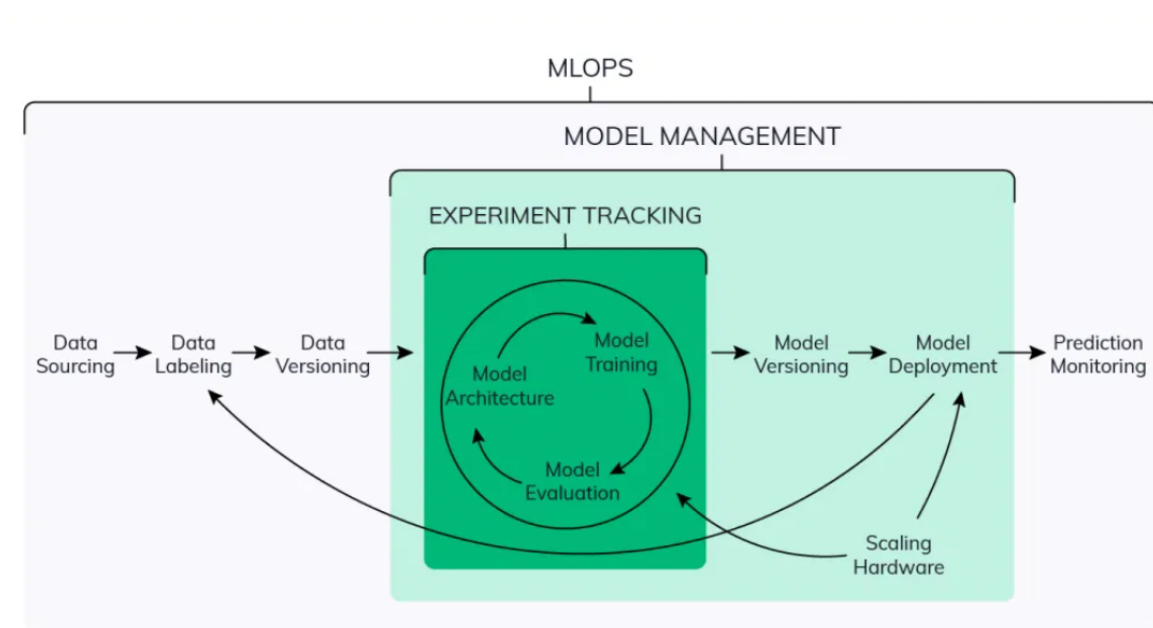


Figure 2.2: MLOps cycle and machine learning experiment tracking

2.3.2 Model Management

Model management is a critical aspect of the machine learning lifecycle, encompassing the processes and tools used to effectively organize, version, and track machine learning models. Effective model management ensures that models can be deployed, monitored, and updated seamlessly, maintaining their performance and relevance over time. The main components of model management:

1. Versioning

- **Importance:** Versioning allows data scientists to keep track of different iterations of a model, ensuring that they can reproduce results and compare performance across versions.
- **Tools:** Tools like MLflow, DVC, and Git provide functionalities to version models, track changes, and manage model metadata.

2. Deployment

- **Purpose:** Deployment involves taking a trained model and making it available for inference in a production environment.
- **Methods:** Models can be deployed as a python function, in a docker container, as an APIs, embedded in applications, as a batch job in Apache Spark or integrated into data pipelines. Platforms like Kubernetes, AWS SageMaker, and Azure ML facilitate seamless deployment.

3. Monitoring

- **Need:** Continuous monitoring of models in production is essential to ensure they perform as expected and to detect any performance degradation or data drift.
- **Metrics:** Key metrics to monitor include accuracy, latency, throughput, and error rates. Tools like Prometheus, Grafana, and custom logging solutions can be used for monitoring.

4. Retraining

- **Process:** As new data becomes available, models may need to be retrained to maintain their performance.
- **Automation:** Automated retraining pipelines can be set up to periodically retrain models, incorporating the latest data and ensuring that the model remains up-to-date.

5. Governance and Compliance

- **Compliance:** Ensuring that models comply with regulatory standards and organizational policies is crucial, especially in industries like finance and healthcare.
- **Documentation:** Proper documentation and audit trails are necessary for compliance and to provide transparency into how models are developed and used.

6. Collaboration

- **Teamwork:** Effective model management facilitates collaboration among data scientists, engineers, and business stakeholders.
- **Tools:** Collaborative tools like Jupyter Notebooks, GitHub, and MLflow allow teams to share code, experiments, and insights.

Benefits of Effective Model Management

- **Reproducibility:** Ensures that models can be consistently reproduced, which is crucial for validation and debugging.
- **Scalability:** Facilitates the scaling of machine learning efforts across an organization.
- **Efficiency:** Streamlines the deployment and monitoring processes, reducing time to market for new models.
- **Compliance:** Helps maintain compliance with legal and regulatory requirements.

In summary, model management is essential for maintaining the lifecycle of machine learning models, from development through deployment and monitoring, ensuring they continue to deliver value and perform optimally in production environments.

2.3.3 Model Registry in Machine Learning

WHY MODEL REGISTRY

Model registry is an essential tool for managing the lifecycle of machine learning models, ensuring that they are well-organized, reproducible, and efficiently deployable, while enhancing collaboration and compliance within an organization.

A model registry is a centralized repository that stores and manages machine learning models. It plays a crucial role in the machine learning lifecycle by providing a systematic way to organize, version, and track models. The key components and benefits of a model registry:

1. Versioning

- **Purpose:** Keeps track of different iterations and versions of a model.
- **Functionality:** Allows comparison of model performance over time and ensures reproducibility.

2. Metadata Management

- **Purpose:** Stores important information about models, such as hyperparameters, training data, and evaluation metrics.
- **Functionality:** Facilitates model auditability and governance.

3. Lifecycle Management

- **Purpose:** Manages the lifecycle of models from development to deployment and monitoring.
- **Functionality:** Ensures smooth transitions between different stages of the model lifecycle.

4. Access Control

- **Purpose:** Defines who can access and modify models in the registry.
- **Functionality:** Enhances security and ensures that only authorized users can make changes.

5. Deployment Integration

- **Purpose:** Facilitates the deployment of models to production environments.
- **Functionality:** Provides integration with deployment tools and platforms for seamless model serving.

As we continue to generate more models, we need a model registry consisting of a tracking server to keep track of the models. Once you've decided that some of the model are ready for production, you then register the model into the mlflow model registry. Within the model registry, we have a staging , production and archive area for the models. In the model registry, all the models that are ready for production are stored here.

Model Registry

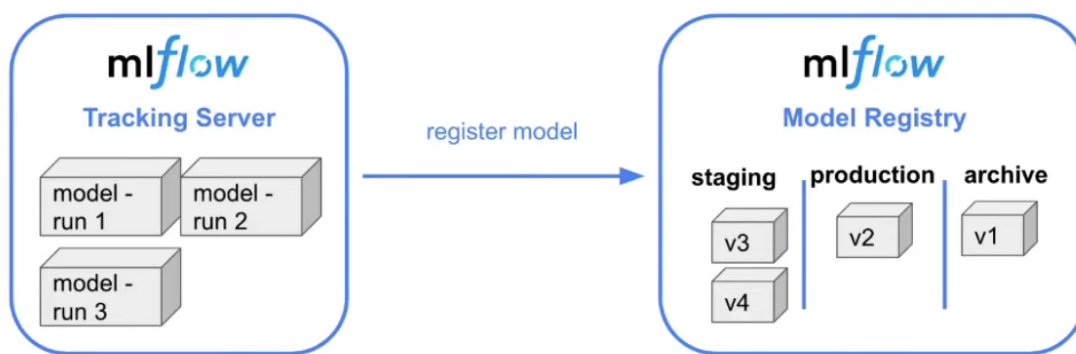


Figure 2.3: Model Registry

The MLOps Engineer can look at the models in the registry, inspect their hyperparameters, the size of the model and so on. He can then decide to move the models between the different stages. The model that is ready for production can be assigned to “staging”, the current model that is used in production can be assigned to the “production” stage, and some of the models can be archived in the “archived” stage. The archived models can be retrieved from the archived stage and move them back to production if we want to roll back some deployments. The model registry is not deploying any model, it just lists the models that are production ready and the stages are just labels assigned to the model. The model registry will need to be complemented with some CI/CD code in order to do the actual deployment of the models.

USING MLFLOWCLIENT

`mlflow.client` is a module in MLflow that provides a way to interact programmatically with an MLflow tracking server. The primary class in this module is `MlflowClient`, which provides a comprehensive API for managing experiments, managing runs, models, logging artifacts, and interacting with model registry.

2.3.4 Interacting Programmatically with MLflowClient

The `mlflow.client` module is useful for advanced use cases where you need programmatic control over MLflow entities, especially in production environments where automated workflows and integrations are required.

Here's a basic example demonstrating how to use `MlflowClient`:

```

1 import mlflow
2 from mlflow.tracking import MlflowClient
3
4 # Initialize the client
5 client = MlflowClient()
6
7 # Create an experiment
8 experiment_id = client.create_experiment("My New Experiment")
9
10 # Start a new run in the experiment
11 run = client.create_run(experiment_id)
12 run_id = run.info.run_id
13
14 # Log parameters, metrics, and tags
15 client.log_param(run_id, "param1", 5)
16 client.log_metric(run_id, "metric1", 0.89)
17 client.set_tag(run_id, "tag1", "value1")
18
19 # Log an artifact (a file)
20 with open("output.txt", "w") as f:
21     f.write("Hello, world!")
22 client.log_artifact(run_id, "output.txt")
23
24 # End the run
25 client.set_terminated(run_id)
26
27 # Get information about the run
28 run_info = client.get_run(run_id)
29 print(run_info)

```

Listing 2.1: Example Usage of `MlflowClient`

Key Methods of `MlflowClient`

1. Experiment Management

- `create_experiment(name, artifact_location=None, tags=None)`: Creates a new experiment.
- `get_experiment(experiment_id)`: Retrieves an experiment by ID.
- `delete_experiment(experiment_id)`: Deletes an experiment.

2. Run Management

- `create_run(experiment_id, start_time=None, tags=None)`: Starts a new run.
- `log_param(run_id, key, value)`: Logs a parameter.
- `log_metric(run_id, key, value, timestamp=None, step=None)`: Logs a metric.
- `set_tag(run_id, key, value)`: Sets a tag.
- `log_artifact(run_id, local_path, artifact_path=None)`: Logs an artifact.

3. Model Registry Management

- `create_registered_model(name)`: Creates a new registered model.
- `create_model_version(name, source, run_id, tags=None, run_link=None, description=None)`: Creates a new model version.
- `transition_model_version_stage(name, version, stage)`: Transitions a model version to a new stage.

2.4 MLflow in Practice

2.4.1 Different scenarios for running MLflow

Let's consider these three scenarios:

- **A single data scientist participating in an ML competition:** In this scenario, having remote tracking server will be an overkill. Saving this information locally will be enough. Also, using model registry is useless since the data scientist is not deploying this model to production.
- **A cross-functional team with one data scientist working on an ML model:** Here, sharing the experiment information is important. Also, using model registry will be a good idea but it can be run remotely or local host.
- **Multiple data scientists working on multiple ML models:** Since multiple data scientist are working on multiple model, collaboration and sharing experiment information is very important. One data scientist can build the model, another data scientist can tune different hyperparameters to add to the model. They need a way to keep track of the models using a remote tracking server. Also, it is important to manage the lifecycle of the model since multiple people build and deploy the model hence model registry is important.

2.4.2 Things to Consider before Configuring MLflow

There are different things to consider before configuring MLflow.

- **Backend Store:** Where MLflow saves information about your experiment such as metadata, models etc
 - ➡ local filesystem
 - ➡ SQLAlchemy compatible DB (e.g. SQLite)
- **Artifacts Store:** Decide where to store the artifact i.e., locally or remote
 - ➡ local filesystem
 - ➡ remote (e.g. s3 bucket)
- **Tracking Server:** Decide how to run tracking server
 - ➡ no tracking server
 - ➡ localhost
 - ➡ remote

2.5 MLflow: Benefits, limitations and alternatives

2.5.1 Remote tracking server

The tracking server can be easily deployed to the cloud. Some of the benefits are:

- Shared experiments with other data scientist
- Collaborate with others to build and deploy models
- Give more visibility of the data science efforts

2.5.2 Issues with running a remote (shared) MLflow server

Some issues can arise when running a remote MLflow server. They include:

- Security
 - ➡ Restrict access to the server (e.g. access through VPN)
- Scalability
 - ➡ Check [Deploy MLflow on AWS Fargate](#)
 - ➡ Check [MLflow at Company Scale](#)
- Isolation
 - ➡ Define standard for naming experiments, models, and a set of default tags
 - ➡ Restrict access to artifacts (e.g. use s3 buckets living in different AWS accounts)

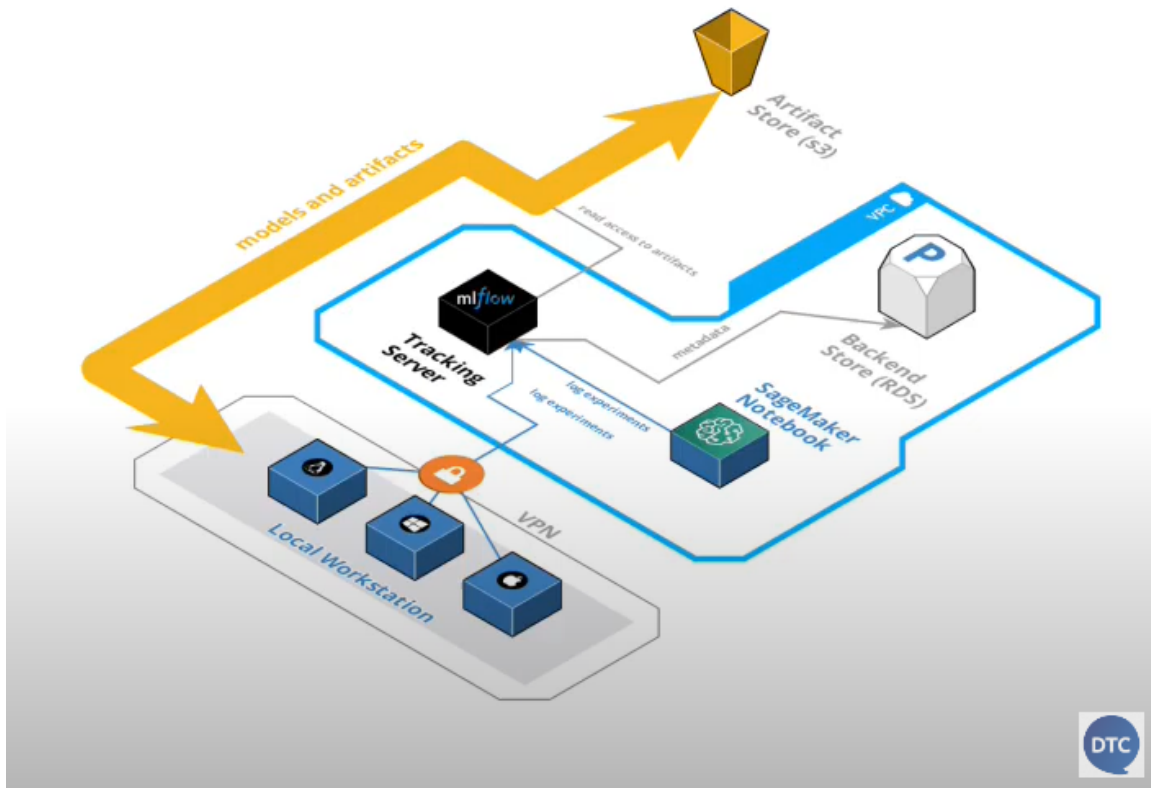


Figure 2.4: Remote Tracking Server with MLflow

2.5.3 MLflow limitations (and when not to use it)

There are some limitations for MLflow. For example:

- **Authentication & Users:** The open source version of MLflow doesn't provide any sort of authentication.
- **Data versioning:** To ensure full reproducibility, we need to version the data used to train the model. MLflow doesn't provide a built-in solution for that but there are a few ways to deal with this limitation.
- **Model/Data Monitoring & Alerting:** This is outside the scope of MLflow and currently there are more suitable tools for doing this.

There are some paid alternatives to MLflow including Neptune.ai, Comet, Weight & Biases and many more.

ORCHESTRATION AND ML PIPELINES

3.1 Introduction: ML pipelines and Mage

3.1.1 Operationalizing ML models

FUN FACT: Operationalizing ML models

Operationalizing ML models involves moving them from development to production to drive business value

To operationalize ML models, we take the following steps:

- **Preparing the model for deployment** involves optimizing performance, ensuring it handles real-world data, and packaging it for integration into existing systems.
- **Deploying the model** involves moving it from development to production, making it accessible to users and applications.
- Once deployed, models must be **continuously monitored for accuracy and reliability**, and may need retraining on new data and updates to maintain effectiveness.
- **Integrating operationalized model into existing workflows, applications, and decision-making processes** to drive business impact.

Effective operationalization enables organizations to move beyond experimentation and drive tangible value from ML at scale, powering intelligent applications that personalize the customer experience and creates real business value.

3.1.2 Why we need to operationalize ML

- **Productivity** - MLOps fosters collaboration between data scientists, ML engineers, and DevOps teams by providing a unified environment for experiment tracking, feature engineering, model management, and deployment. This breaks down silos and accelerates the entire machine learning lifecycle.

- **Reliability** - MLOps ensures high-quality, reliable models in production through clean datasets, proper testing, validation, CI/CD practices, monitoring, and governance.
- **Reproducibility** - MLOps enables reproducibility and compliance by versioning datasets, code, and models, providing transparency and auditability to ensure adherence to policies and regulations.
- **Time-to-value** - MLOps streamlines the ML lifecycle, enabling organizations to successfully deploy more projects to production and derive tangible business value and ROI from AI/ML investments at scale.

3.1.3 How Mage helps MLOps

- **Data preparation** - Mage offers features to build, run, and manage data pipelines for data transformation and integration, including pipeline orchestration, notebook environments, data integrations, and streaming pipelines for real-time data.
- **Training and deployment** - Mage helps prepare data, train machine learning models, and deploy them with accessible API endpoints.
- **Standardize complex processes** - Mage simplifies MLOps by providing a unified platform for data pipelining, model development, deployment, versioning, CI/CD, and maintenance, allowing developers to focus on model creation while improving efficiency and collaboration.

3.1.4 Project setup for Mage

Clone the following repository containing the complete code for this module:

```
git clone https://github.com/mage-ai/mlops.git
```

Change directory into the cloned repo:

```
cd mlops
```

Launch Mage and the database service (PostgreSQL):

```
./scripts/start.sh
```

If you don't have bash in your environment, modify the following command and run it:

```
PROJECT_NAME=mlops \  
  MAGE_CODE_PATH=/home/src \  
  SMTP_EMAIL=$SMTP_EMAIL \  
  SMTP_PASSWORD=$SMTP_PASSWORD \  
  docker compose up
```

The subproject that contains all the pipelines and code is named `unit_3_observability`

3.2 Data preparation: ETL and Feature Engineering

3.3 Training: sklearn models and XGBoost

3.4 Observability: Monitoring and Alerting

3.5 Triggering: Inference and Retraining

3.6 Deploying: Running operations in Production

DEPLOYMENT

4.1 Model Deployment

4.1.1 Three ways of deploying models

FUN FACT: Deploying ML models

Deploying machine learning model models is an iterative process that often involves multiple rounds of training, testing, and validating before the model is ready for production.

Deploying a model means that other application can get predictions from our model. There are three modes of deployment namely **online** deployment, **offline** or batch deployment, and **streaming**.

First, we need to ask ourselves if we need to have predictions immediately or it can wait a little bit for an hour, a day or a week. If it can wait for a little bit, then we go for batch deployment. i.e., the model is not running all the time and we just apply our model to new data regularly. In online deployment, the model is up and running all the time and is always available. Two variant of online deployment is deployment as a web service and deployment via streaming. In web service deployment, we send HTTP requests and the service sends out prediction. In streaming, there is an “events model service” listening for events on the stream and reacting to this event.

4.1.2 Web services: Introduction to Flask

WEB SERVICE

A web service is a method for communicating between two devices over a network using some protocols.

Assuming we want to use our model inside a `churn` service in order to make some predictions. The `marketing` service will communicate with our `churn` service by

sending some request and getting a response. This can be done using a web service. In **web service**, a user sends a request in the form of a query, then the web service sends back a response to the user with the result. In **Fig. 4.1**, we have a notebook that was used to train the churn model. The model is saved to file. We can load this model from a different process or web service called the churn service. If the marketing service wants to identify if the user will churn, they send a request to the churn service with information about the user then they get back the predictions and based on these predictions, the marketing service can decide whether they want to send a promotional email with say 25% discount to prevent churn.

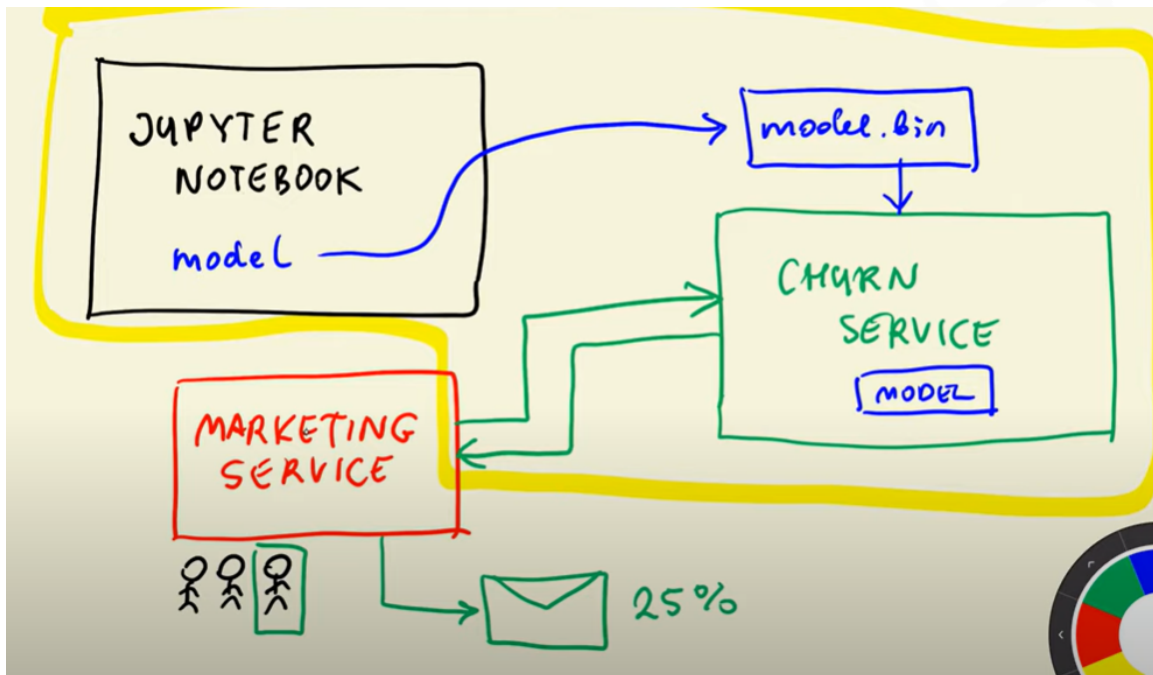


Figure 4.1: Serving Churn Model

To do this, we put the model inside a web service using **flask**, which is a framework for creating web services in python, then we isolate dependencies for this service so they don't interfere with other services on ur machine by creating a special environment for python dependencies using **Pipenv**. Then we add another layer with system dependencies using **Docker**, and then finally we deploy the container containing this model to the cloud using **AWS Elastic Beanstalk**.

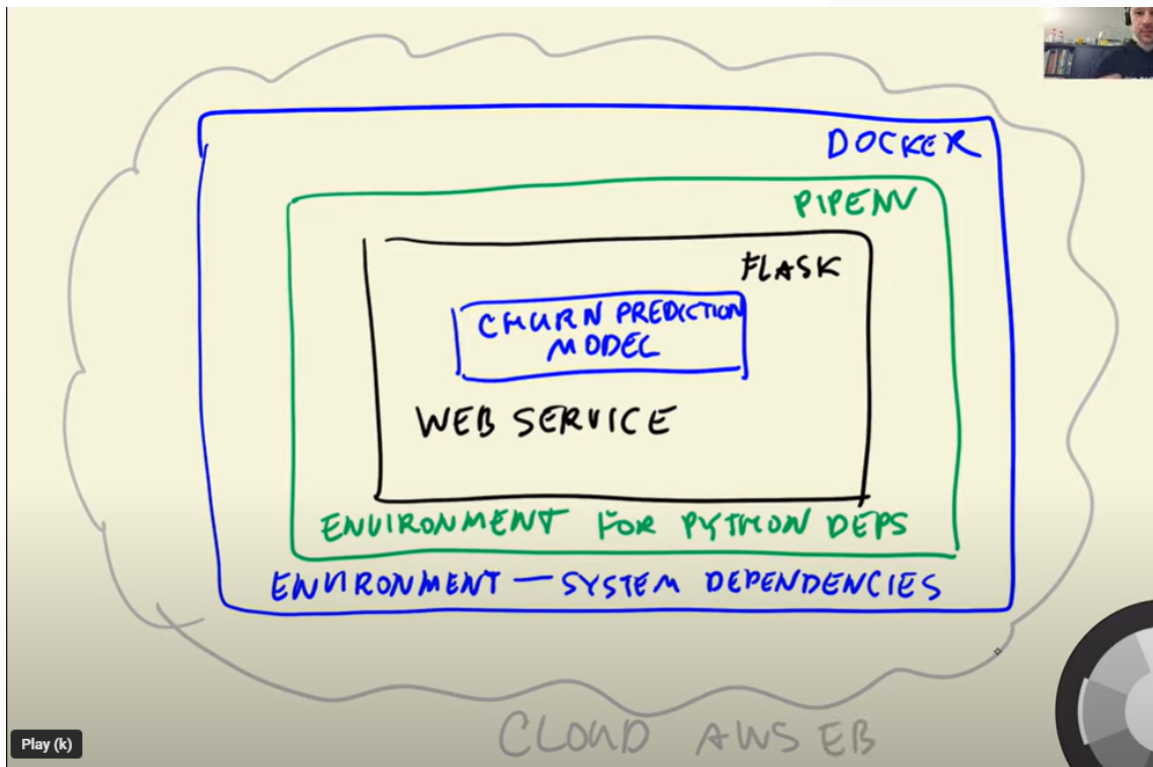


Figure 4.2: Layers to deployment

To deploy a model, we turn our notebook to Python script called `train.py` where we save the model as a pickle file. Then using our `predict.py`, we can load the model and make prediction.

There are some methods in web services we can use it to satisfy our problems. Here below we would list some.

- GET: GET is a method used to retrieve files, For example when we are searching for a cat image in google we are actually requesting cat images with GET method.
- POST: POST is the second common method used in web services. For example in a sign up process, when we are submitting our name, username, passwords, etc we are posting our data to a server that is using the web service. (Note that there is no specification where the data goes)
- PUT: PUT is same as POST but we are specifying where the data is going to.
- DELETE: DELETE is a method that is used to request to delete some data from the server.

We can create a simple service using flask that pings and send a response back. To do that, we create a `ping.py` file containing:

```
from flask import Flask
```

MACHINE LEARNING OPERATIONS

```
app = Flask('ping')

@app.route('/ping', methods=['GET'])
def ping():
    return "PONG"

if __name__ == '__main__':
    app.run(debug=True, host='0.0.0.0', port=9696)
```

We start by installing and importing flask

```
from flask import Flask
```

We create a flask app

```
app = Flask('ping')
```

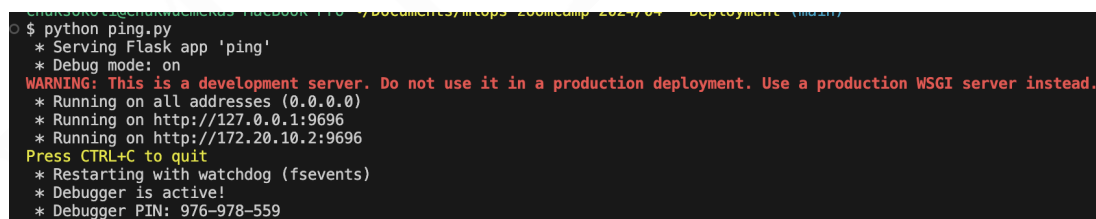
and then add a decorator, which is a way to add some extra functionality to our functions. This extra functionality will allow us turn the function into a web service. We specify the address the 'ping' will live in, and the method to access this route.

```
@app.route('/ping', methods=['GET'])
```

We run the app in debug mode and specify the host to run on.

```
if __name__ == '__main__':
    app.run(debug=True, host='0.0.0.0', port=9696)
```

We get the following output when it runs:



```
o $ python ping.py
* Serving Flask app 'ping'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:9696
* Running on http://172.20.10.2:9696
Press CTRL+C to quit
* Restarting with watchdog (fsevents)
* Debugger is active!
* Debugger PIN: 976-978-559
```

Figure 4.3: Ping result

To test it, open your browser and search localhost:9696/ping, You'll see that the 'PONG' string is received. Congrats You've made a simple web server.

4.1.3 Serving the Churn Model with Flask

In serving the churn model with flask, we want the model to be available at /predict in the churn service. The marketing service will send the churn service with information about the customers, and then we reply them with the probability of churning. The churn service also sends a promotional email to the customer with 25% discount. To make the web service predict the

churn value for each customer, we need to first load the previous saved model and use a prediction function in a special route.

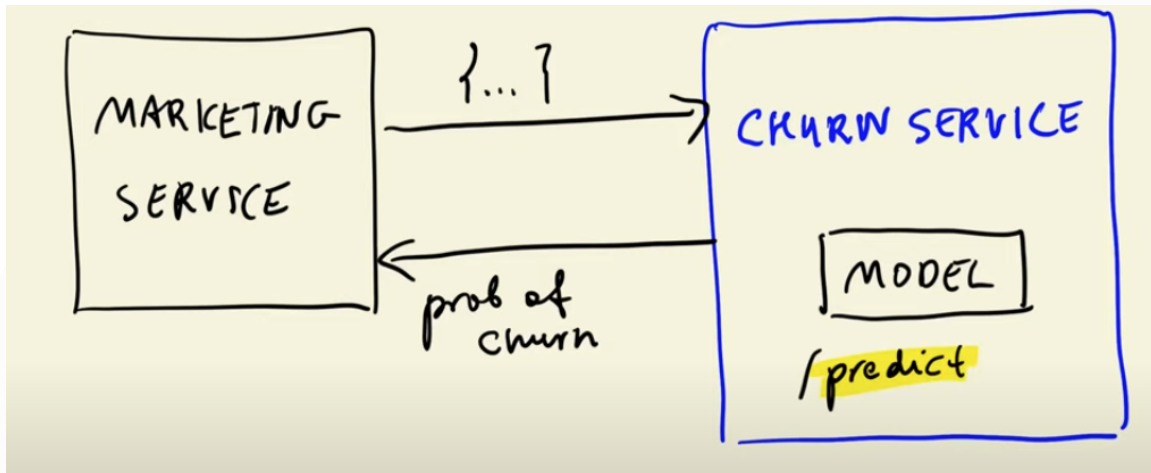


Figure 4.4: Ping result

- To load the previous saved model, we create a `predict.py` file that loads the model

```
# Load the model
import pickle
from flask import Flask
from flask import request
from flask import jsonify

model_file = 'model_C=1.0.bin'

# load the model
with open(model_file, 'rb') as f_in: # read binary file
    dv, model = pickle.load(f_in)    # load() loads the file
```

- A function to predict a value for a customer

```
def predict_single(customer, dv, model):
    X = dv.transform([customer]) # one-hot encoding to the data
    y_pred = model.predict_proba(X)[:, 1]
    return y_pred[0]
```

- Then the function used to create the web service that makes prediction and sends an email to the customer.

```
app = Flask('churn') # name of the app
@app.route('/predict', methods=['POST']) # in order to send the
    customer information we need to post its data.
def predict():
    customer = request.get_json() # web services work best with
    json frame
    prediction = predict_single(customer, dv, model)
```

MACHINE LEARNING OPERATIONS

```
churn = prediction >= 0.5

result = {
    'churn_probability': float(prediction), # cast numpy float
    type to python native float type
    'churn': bool(churn), # casting the value using bool
    method
}
return jsonify(result) # send back the data in json format to
the user
```

To deploy, the customer information in the `deploy.py` file is sent as JSON. The predict script is turned into a web service and sends a response back to the marketing service with predictions as JSON and an email to the customer likely to churn.

```
customer_id = "asdx-123d"
customer_email = "asdx-123d@yahoo.com"
customer = {
    "gender": "female",
    "seniorcitizen": 0,
    "partner": "yes",
    "dependents": "no",
    "phoneservice": "no",
    "multiplelines": "no_phone_service",
    "internetservice": "dsl",
    "onlinesecurity": "no",
    "onlinebackup": "yes",
    "deviceprotection": "no",
    "techsupport": "no",
    "streamingtv": "no",
    "streamingmovies": "no",
    "contract": "two_year",
    "paperlessbilling": "yes",
    "paymentmethod": "electronic_check",
    "tenure": 10,
    "monthlycharges": 29.85,
    "totalcharges": (2 * 29.85)
}
# Making requests
import requests
url = "http://localhost:9696/predict"
response = requests.post(url, json=customer).json()
print(response)

if response["churn"]:
    print(f"Sending email to {customer_id} with email:", {customer_email
    })
else:
    print(f"Customer {customer_id} will not churn")
```

When you run your app you will see a warning that it is not a WSGI server and not suitable for production environments. To fix this issue and run this as a production server there are plenty of ways available. One way to create a WSGI server is to use gunicorn. To install it use the command

```
pip install gunicorn
```

And to run the WSGI server you can simply run it with the command

```
gunicorn --bind 0.0.0.0:9696 churn:app.
```

Note that in `churn:app` the name `churn` is the name we set for our the file containing the code `app = Flask('churn')` (for example: `churn.py`). You may need to change it to whatever you named your Flask app file. So far, we have been able to make a production server that predict the churn value for new customers.

4.1.4 Dependencies and Environment Management: Pipenv

4.2 Online Deployment

4.2.1 Web services: Deploying models with Flask and Docker

- Like this one,

which is wrapped in gray. I use it for notes...

- Or this one,

which is wrapped in red. I use it for fun facts or other asides...

- Or this one,

which is wrapped in blue and used for mathy stuff.

- Or this last one,

which is wrapped in green. With a title, it's used for enumerated examples (see `\extitle` and `\excounter`). Observe:

EXAMPLE 4.1: Test

This is an example. What's the answer to $2 + 2$?

ANSWER: Obviously 4, lol.

EXAMPLE 4.2: Test Again

This one will increment the counter automatically, resetting for each chapter.

- For red and blue boxes, there are custom commands for titles, too:

ONE TITLE

Like this

TWO TITLES: A Subtitle

Or this

These styles also automatically apply to theorems and claims.

Theorem 4.1 (Pythagorean Theorem). *For any right triangle with legs a, b and hypotenuse c :*

$$a^2 + b^2 = c^2 \quad (4.1)$$

Proof. This is left as an exercise to the reader. ■

Claim 4.1. *This is the greatest note template in the world.*

There are different ways to quote things, too, depending on how you want to emphasize:

This is a simple, indented quote with small letters and italics usually suitable for in-text quotations when you just want a block.

Alternatively, you can use the `\inspiration` command from the chapter heading, which leverages the `thickleftborder` frame internally, but adds a little more padding

and styling (there’s also just `leftborder` for a thinner variant):

■ Hello there!

4.3 On Cross-Referencing

You can reference most things—see [Theorem 4.1](#) or [\(4.1\)](#) or the [Introduction](#) chapter—directly and easily as long as you give them labels. These are “built-ins.” However, you can also create a [custom term](#) that will be included in the index, then include references to it that link back to the original definition. Try clicking: [custom term](#). Building the index is on you, though. You can also reference by using a different term for the text: [like this](#). Sometimes it doesn’t fit the [grammatical structure](#) of the sentence so you can define the term one way and visualize it another way (this creates a [grammar](#) entry in the index). There’s also [math terms](#) and a way to reference them: [math terms](#) (clickable), but they do **not** show up in the index.

4.4 On Math

Most of the math stuff is just macros for specific things like the convolution operator, \otimes , probabilities, $\Pr[A|B = C]$, or big- O notation, $\mathcal{O}(n^2 \log n)$ but there’s also a convenient way to include explanations on the side of an equation:

$$\begin{array}{ll} 1 + 1 \stackrel{?}{=} 2 & \text{first we do this} \\ 2 \stackrel{?}{=} 2 & \text{then we do this} \\ 2 = 2 & \blacksquare \end{array}$$

These are all in the `CustomCommands.sty` file.