Week 2: Experiment Tracking

1. Experiment Tracking
2. Experiment tracking is the process of keeping track of all the relevant information from an ML experiment
   1. Source code
   2. Environment
   3. Data
   4. Model
   5. Hyperparameters
   6. Metrics
3. Why is tracking experiments in spreadsheets not enough?
   1. Error prone
   2. No standard format
   3. Visibility and collaboration
4. ML Flow contains four main modules:
   1. Tracking
   2. Models
   3. Model Registry
   4. Projects
5. Tracking experiments with MLFlow
   1. The MLFlow Tracking module allows you to organize your experiments into duns and to keep track of:
      1. Parameters
      2. Metrics
      3. Artifacts
      4. Modles
   2. Along with this information, MLFlow automatically logs extra information about the run:
      1. Source code
      2. Version of the code (git commit)
      3. Start and end time
      4. Author

A diagram of a lifecycle

AI-generated content may be incorrect.

1. MLFlow: Benefits, Limitations, and Alternatives
2. Remote tracking server
   1. The tracking server can be easily deployed to the cloud
   2. Some benefits:
      1. Share experiments with other data scientists
      2. Collaborate with others to build and deploy models
      3. Give more visibility of the data science efforts
3. Issues with running a remote (shared) MLFLow server
   1. Security: Restrict access to the server (eg access through VPN)
   2. Scalability:
      1. Check Deploy MLFlow on AWS Fargate
      2. Check MLFlow at Company Scale by Jean-Denis Lesage
   3. Isolation:
      1. Define standard for naming experiments, models, and a set of default tags
      2. Restrict access to artifacts (eg use S3 buckets living in different AWS accounts)
4. MLFlow Limitations (and when not to use it)
   1. Authentication & Users: The open source version of MLFlow doesn’t provide any sort of authentication
   2. 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.
   3. Model/Data Monitoring & Alerting: This is outside the scope of MLFlow and currently there are more suitable tools for doing this.
5. MLFlow Alternatives:
   1. Neptune
   2. Comet
   3. Weights & Biases
   4. 15 Best Tools for ML Experiment Tracking and Management