Data Governance and Lineage Toolkit for Financial ML Models

Introduction

In financial machine learning systems, data governance and lineage tracking are critical to ensure transparency, reproducibility, and regulatory compliance. Regulations such as GDPR (General Data Protection Regulation) and BCBS 239 (Basel Committee on Banking Supervision's Principles for effective risk data aggregation and risk reporting) require organizations to maintain detailed records of data sources, transformations, and model versions.

Importance of Data Lineage

- **Traceability:** Ability to trace data from origin to final model predictions, essential for auditing.
- **Provenance:** Information about the source and modifications to datasets help prove data integrity.
- **Versioning:** Tracking changes in data and models supports reproducibility and rollback in case of errors or policy violations.

Workflow for Logging and Lineage Tracking

We propose the following workflow:

1. **Data Ingestion:** Record metadata such as data source (URL, database, vendor), collection time, schema, and a cryptographic hash to verify integrity.
2. **Data Transformation:** Log all preprocessing steps, including feature engineering and cleaning. Record transformation scripts or pipeline versions (e.g., using git commit hashes).
3. **Model Versioning:** Link each training run to a specific dataset version and code version (e.g., git commit or container image ID).

4. **Evaluation and Deployment:** Store results, model artifacts, and logs in a secure database or append-only storage to comply with audit trails.

5. **Monitoring and Feedback:** Continuously monitor model performance and data drift; log any feedback or manual interventions.

Regulatory Compliance

Proper data lineage ensures compliance with:

- **GDPR:** Right to explanation and data erasure require detailed data and model records.
- **BCBS 239:** Emphasizes accuracy, completeness, and auditability of risk data.
- **SOX:** Sarbanes-Oxley Act requires data integrity and transparent reporting in financial institutions.

By implementing comprehensive data and model lineage logging, financial institutions can mitigate risks, enable faster audits, and improve trust in AI-driven decisions.