Loan Risk Prediction Machine Learning Model Pipeline
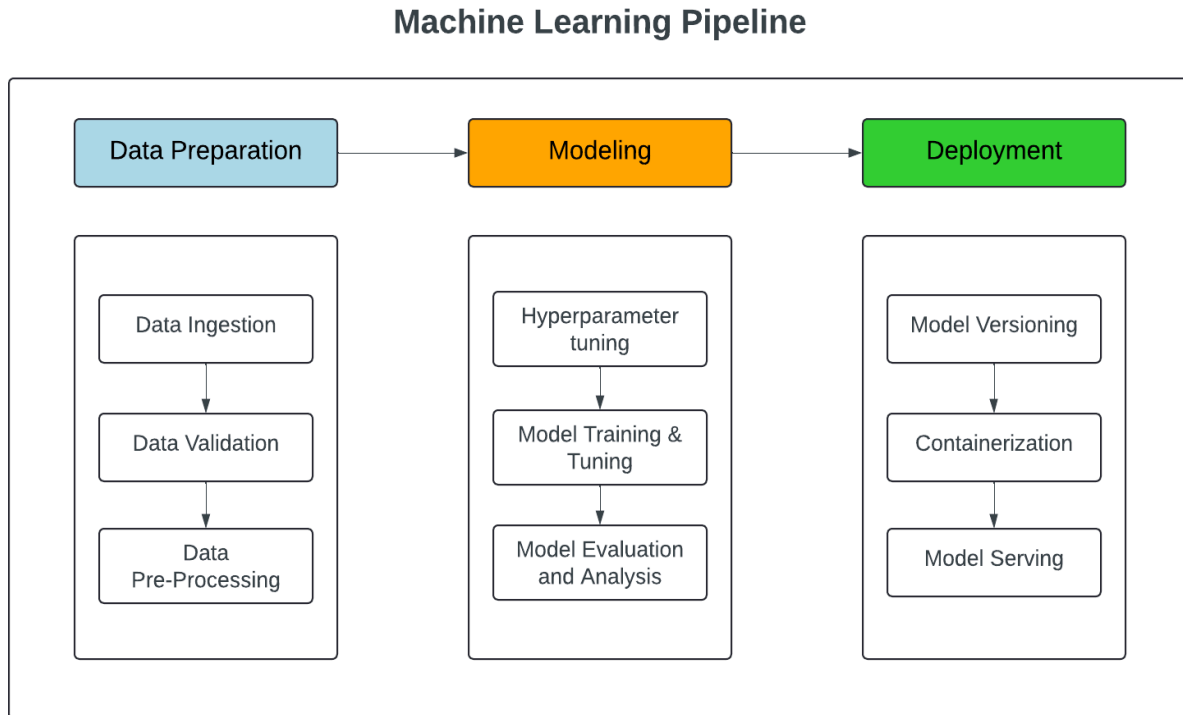
Machine Learning Pipeline

1. Data Preparation

1. **Data Ingestion:** Loan data from CSV files (loan, payment, and clarity underwriting) is ingested. Integration with databases and APIs can be implemented for future expansion.

2. **Data Validation:** The dataset undergoes validation to ensure conformity with the required schema defined in schema.json.

3. **Data Pre-Processing:**
   - Handling missing values by dropping the columns or the rows with high missing value or filling with mean, median or mode.
   - Remote highly correlated columns, drop duplicates and data type conversion of number and category.
   - A train-test split (80/20) is established for model training.

2. Modeling

1. **Hyperparameter Tuning:**

   ○ **FLAML AutoML** automates the tuning process for LightGBM hyperparameters, optimizing for accuracy. This provides a faster and more efficient approach.

2. **Model Training & Tuning:**

   ○ The model is trained using the optimized hyperparameters identified by FLAML AutoML.

   ○ Training metrics such as accuracy and loss are logged for continuous monitoring.

3. **Model Evaluation and Analysis:**

   ○ Model performance is evaluated using metrics like **accuracy**, **precision**, and **recall**.

   ○ Feature importance is visualized to understand the factors influencing model decision-making.

   ○ **MLflow** is utilized to log and track model versions, hyperparameters, and performance metrics, ensuring easy reproducibility.

3. Deployment

1. **Model Versioning:** Each iteration of the model is versioned for clear tracking of updates.
2. **Containerization:** The model and its dependencies are packaged into a Docker container for consistent deployment.
3. **Model Serving:** The Docker image can be deployed on AWS SageMaker, Databricks, or GCP. I've set it up to upload the image to Render.com, where the model serves loan application predictions via an API.

Monthly retraining addresses data drift, with MLflow for logging and versioning. AutoML automates tuning, and GitHub Action triggers Docker deployment each month.