Assignment 2
Machine Learning Pipelines

CS611 - Machine Learning Engineering
Version: April 2025

Objectives

The objectives of this exercise are as follows:

1. Build Machine Learning (ML) end-to-end pipelines that
    a. trains ML models and store the model artefacts in a model bank,
    b. retrieves model, makes predictions / inference and
    c. monitor model performance and stability across time.
2. Work with Docker containers to manage environments and dependencies
3. Practice deck building, documentation and presentations

Context

You are a data scientist working at a financial institute (e.g. bank). Your company lends money to users in the form of cash loans. You are tasked to build a machine learning model that can predict whether a user will default on their loan at the point of loan application.

For this assignment, you will be building and serving ML models through production ML pipelines. You have 2 tasks: build an end-to-end ML pipeline, and prepare a presentation deck (max 10 slides) to present to your manager, other engineers and business users of what your ML pipeline contains.

Task 1: Build end-to-end ML pipeline (10 marks)

You are to build an end-to-end ML pipeline that performs the following:

- Train ML models, evaluates and selects the best model to be stored in a model store (could be just a folder of artefacts for this assignment)
- Retrieves best models to make predictions / inference across a time period and stores the model predictions as a gold table in the datamart
- Monitor the performance and stability of the model predictions across a time period, stores the monitoring results as a gold table in the datamart
- Visualise the performance and stability of the model across a time period
- Decide on model governance and SOP on when / how to refresh the ML model in production, and deployment options.

Alongside your work that you did to prepare your datamart from Assignment 1, you are provided with 3 sample completed Assignment 1 datamart data pipelines by your classmates. This simulates other data sources from other tech-families within your company. You may explore the work and documentations of these other datamarts and decide which features you want to combine with yours or to use theirs entirely; the decision is up to you! Most importantly is to explain your design choice in task 2 (presentation deck).

You are encouraged to use open-source tooling such as Airflow to orchestrate and schedule your pipeline. Your pipeline should be able to be backfilled to show what happens across time. Refer to Lab 5 on how to build a DAG with Airflow and how to trigger a backfill.

Task 2: Build a presentation deck that documents your ML pipeline (5 marks)

Build a presentation deck on your design choices and implementation from Task 1. Assume that this deck will be circulated to technical and non-technical colleagues and that you do not have a chance to present it to them; build the deck as a slideument.

A **slideument** is a hybrid between a **slide presentation** and a **document** — a presentation slide that is designed to function as both a visual aid during a presentation and a handout document. (very common in the business world!)

This is a powerful exercise as it allows you to sharpen your deck building, visualisation and storytelling skills which are vital at the workplace for visibility of your work. You can also get to use this in future tech interviews if ever they ask you what experience you have on machine learning engineering! The real benefactor of this task is really you! :)

Data Provided

You are provided with 3 sample completed Assignment 1 data pipelines by your classmates. You can use these completed data pipelines alongside your data pipeline from Assignment 1.

You are provided with the following data about users:

1. feature_clickstream.csv
2. feature_attributes.csv
3. feature_financials.csv

You are provided with the following data about loans:

1. lms_loan_daily.csv

You are encouraged to perform Exploratory Data Analytics to discover what features you want to build in your pipeline.

BE CAREFUL OF DATA LEAKAGE! Data leakage (also called leakage or target leakage) is a common issue in machine learning where information from outside the training dataset — specifically from the future or from the target variable — is accidentally used to create the model. This leads to overly optimistic performance during training or validation, but poor generalization to unseen data. Types of Data Leakage:

- Target Leakage: Happens when the model has access to data that would not be available at prediction time. Example: Including a column like "Loan Paid Off" when trying to predict loan default — that’s the answer!
- Train-Test Contamination: Occurs when the test data somehow influences the training process. Example: Normalizing your full dataset before splitting into train/test, so info from test leaks into training.
- Temporal Leakage: Using future data to predict past or present. Example: Using stock prices from a week ahead to predict today’s price.

Submissions

Due date: **25 June 2025**

You are expected to upload the following to eLearn:

1. A zip file of your code artefacts from task 1. The zip folder should unzip to contain the following files and subfolders:
    a. A working Airflow pipeline that runs the whole pipeline.
    b. Dockerfile
    c. docker-compose.yaml
    d. requirements.txt
    e. utils (folder, if you want to put any code in here)
    f. data
    g. datamart
    h. Readme.txt (with just 1 line to your github repo link)
2. PDF presentation deck (max 10 slides) from task 2.

Assessments

We will assess as follows **(total 15 marks)**:

1. We will unzip your zip.file into a folder and run the command in the terminal “**docker-compose buil**d” and “**docker-compose up**”. Your docker-compose up should provide a link to Airflow (similar to Lab 5). If we can enter Airflow from docker-compose up, you get **5 marks**!

2. We will run your Airflow DAG. If your Airflow DAG can run and create this your ML model artefacts, make predictions and monitor your model performance and stability, you get **5 marks**!
3. We will read through your presentation deck and grade it as follows:
    a. ML pipeline technical design decisions and explanations WITH MODEL MONITORING VISUALISATION results: **3 marks**
    b. Is the deck pretty? Is it worthy of corporate / business standards? **2 marks**