Machine Learning Task for TDT4173
Append Consulting AS x Hydro ASA

Task Overview

This task is part of a consultancy project conducted by Append Consulting for Hydro.

Hydro is a Norwegian industrial company with operations in energy, aluminum, and recycling. Append is a consultancy firm mainly consisting of students from NTNU, that works with data science, artificial intelligence, and system development to help companies make better use of data and technology in their operations.

The aim of this task is to develop accurate - but conservative - forecasts of incoming raw material deliveries, to be used in a larger optimization tool developed by Append.

Task Description

You are provided historical data on raw material deliveries and orders through the end of 2024. Each raw material is identified by a unique `rm_id`. The goal is to develop a model that forecasts the cumulative weight of incoming deliveries of each raw material from January 1, 2025, up to any specified end date between January 1 and May 31, 2025.

For any end date within this range, the model should predict the total weight in kg of a raw material delivered from and including January 1, 2025, to and including the end date.

Dataset Overview

The datasets are organized as follows:

- **data/kernel/receivals.csv**: The primary dataset containing historical records of material receivals. Each entry includes a timestamp, the quantity received, and the corresponding `rm_id`.
- **data/kernel/purchase_orders.csv** Contains information on ordered quantities and expected deliveries.
- **data/extended/materials.csv** (Optional): Metadata on various raw materials, including categories and classifications.

- **data/extended/transportation.csv** (Optional): Transportation-related data that could affect delivery times and consistency.

Evaluation

Quantile Error at 0.2 (Asymmetric Loss)

Let there be N raw materials indexed by $i = 1, \ldots, N$. Over a forecasting window of h days, define

$$A_i \;=\; \sum_{t=1}^{h} y_{i,T+t}, \qquad F_i \;=\; \sum_{t=1}^{h} \hat{y}_{i,T+t},$$

as the actual and forecasted total deliveries for material i, respectively.

To evaluate performance, we compute the quantile loss at the 0.2 level:

$$\text{QuantileLoss}_{0.2}(F_i, A_i) \;=\; \max\left(0.2 \cdot (A_i - F_i),\; 0.8 \cdot (F_i - A_i)\right).$$

The overall metric is the average quantile loss across all materials:

$$\text{QuantileError}_{0.2} \;=\; \frac{1}{N} \sum_{i=1}^{N} \text{QuantileLoss}_{0.2}(F_i, A_i).$$

This metric penalizes overestimation more than underestimation, which aligns with the practical needs of smelting. If we underestimate the available materials, the smelt can usually continue with what is on hand. However, if we overestimate, we risk planning a smelt that cannot be completed due to missing resources. Therefore, it's better for the model to be slightly cautious and predict too little rather than too much.