Project Description

The data is stored in three files:

- `gold_recovery_train.csv` — training dataset download
- `gold_recovery_test.csv` — test dataset download
- `gold_recovery_full.csv` — source dataset download

Data is indexed with the date and time of acquisition (`date` feature). Parameters that are next to each other in terms of time are often similar.

Some parameters aren't available, because they were measured and/or calculated much later. That's why some of the features present in the training set may be absent from the test set. There are also no targets in the test set.

The source dataset contains the training and test sets with all of the features.

You have raw data which was downloaded straight from the warehouse. Before creating the model, follow the instructions below to make sure the data is correct.

Project instructions

1. Prepare the data

1.1. Open the files and study the data.

Path to the files:

- */datasets/gold_recovery_train.csv*
- */datasets/gold_recovery_test.csv*
- */datasets/gold_recovery_full.csv*

1.2. Check that recovery was calculated correctly. Using the training set, calculate the recovery for the `rougher.output.recovery` feature. Find the *MAE* between your calculations and the feature values. Describe your findings.

1.3. Analyze the features not available in the test set. What are these parameters? What is their type?

1.4. Perform data preprocessing.

2. Analyze the data

2.1. Take note of how the concentration of metals (*Au, Ag, Pb*) changes depending on the purification stage.

2.2. Compare the feed particle size distributions in the training set and in the test set. If the distributions vary significantly, model evaluation will be performed incorrectly.

2.3. Consider the total concentrations of all substances at different stages of the recovery process: the raw feed, rougher concentrate, and final concentrate. Do you notice any abnormal values in the total distribution? If you do, is it worth removing them from both sets? Describe your findings and eliminate anomalies.

3. Build the model

3.1. Write a function to calculate the final *sMAPE* value.

3.2. Train different models. Evaluate them using cross-validation. Pick the best model and test it using the test set. Describe your findings.

Use these formulas for evaluation metrics:

$$\mathrm{sMAPE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \times 100\%$$

$$\text{Final sMAPE} = 25\% \times \text{sMAPE (rougher)} + 75\% \times \text{sMAPE (final)}$$

Project evaluation

We've put together the evaluation criteria for the project. Read this carefully before moving on to the case.

Here's what the reviewers will look at when reviewing your project:

- Did you correctly prepare and analyze the data?
- Which models did you develop?
- How did you evaluate model quality?
- Did you follow all the steps of the instructions?
- Did you stick to project structure and explain the steps you performed?

- What are your findings?
- Did you keep the code neat and avoid code duplication?

You have your takeaway sheets and the summaries of the previous chapters, so you're all set and good to go.

Good luck!