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 are not available because they were measured and/or calculated much later. That's why, some of the features that are present in the training set may be absent from the test set. The test set also doesn't contain targets.

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

You have the raw data that was only downloaded from the warehouse. Before building the model, check the correctness of the data. For that, use our instructions.

Project instructions

1. Prepare the data

1.1. Open the files and look into the data.

Path to files:

/datasets/gold_recovery_train.csv
/datasets/gold_recovery_test.csv
/datasets/gold_recovery_full.csv

1.2. Check that recovery is calculated correctly. Using the training set, calculate recovery for the `rougher.output.recovery` feature. Find the *MAE* between your calculations and the feature values. Provide 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 concentrations of metals (*Au, Ag, Pb*) change 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, the model evaluation will be incorrect.

2.3. Consider the total concentrations of all substances at different stages: raw feed, rougher concentrate, and final concentrate. Do you notice any abnormal values in the total distribution? If you do, is it worth removing such values from both samples? Describe the 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 sample. Provide findings.

Use these formulas for evaluation metrics:

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

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

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:

- Have you prepared and analyzed the data properly?
- What models have you developed?
- How did you check the model‘s quality?
- Have you followed all the steps of the instructions?
- Did you keep to the project structure and explain the steps performed?
- What are your findings?
- Have you kept the code neat and avoided code duplication?

Remember, the Knowledge Base has everything you need to help you complete this project.

Good luck!