**Постановка задачи**

Подготовьте прототип модели машинного обучения для «Цифры». Компания разрабатывает решения для эффективной работы промышленных предприятий.

Модель должна предсказать коэффициент восстановления золота из золотосодержащей руды. В вашем распоряжении данные с параметрами добычи и очистки.

Модель поможет оптимизировать производство, чтобы не запускать предприятие с убыточными характеристиками.

**Нам нужно:**

1. Подготовить данные;
2. Провести исследовательский анализ данных;
3. Построить и обучить модель.

**Описание данных**

**Технологический процесс**

* *Rougher feed* — исходное сырье
* *Rougher additions* (или *reagent additions*) — флотационные реагенты: *Xanthate, Sulphate, Depressant*
  + *Xanthate* — ксантогенат (промотер, или активатор флотации);
  + *Sulphate* — сульфат (на данном производстве сульфид натрия);
  + *Depressant* — депрессант (силикат натрия).
* *Rougher process* (англ. «грубый процесс») — флотация
* *Rougher tails* — отвальные хвосты
* *Float banks* — флотационная установка
* *Cleaner process* — очистка
* *Rougher Au* — черновой концентрат золота
* *Final Au* — финальный концентрат золота

**Параметры этапов**

* *air amount* — объём воздуха
* *fluid levels* — уровень жидкости
* *feed size* — размер гранул сырья
* *feed rate* — скорость подачи

**Наименование признаков**

Наименование признаков должно быть такое:

[этап].[тип\_параметра].[название\_параметра]

Пример: rougher.input.feed\_ag

Возможные значения для блока [этап]:

* *rougher* — флотация
* *primary\_cleaner* — первичная очистка
* *secondary\_cleaner* — вторичная очистка
* *final* — финальные характеристики

Возможные значения для блока [тип\_параметра]:

* *input* — параметры сырья
* *output* — параметры продукта
* *state* — параметры, характеризующие текущее состояние этапа

*calculation* — расчётные характеристики

**Метрика качества**

Для решения задачи введём новую метрику качества — **sMAPE** (англ. *Symmetric Mean Absolute Percentage Error*, «симметричное среднее абсолютное процентное отклонение»).

Нужно спрогнозировать сразу две величины:

1. эффективность обогащения чернового концентрата rougher.output.recovery;
2. эффективность обогащения финального концентрата final.output.recovery.

**Описание проекта**

Данные находятся в трёх файлах:

* gold\_industry\_train.csv — обучающая выборка;
* gold\_industry\_test.csv — тестовая выборка;
* gold\_industry\_full.csv — исходные данные.

Данные индексируются датой и временем получения информации (признак date). Соседние по времени параметры часто похожи.

Некоторые параметры недоступны, потому что замеряются и/или рассчитываются значительно позже. Из-за этого в тестовой выборке отсутствуют некоторые признаки, которые могут быть в обучающей. Также в тестовом наборе нет целевых признаков.

Исходный датасет содержит обучающую и тестовую выборки со всеми признаками.

В вашем распоряжении сырые данные: их просто выгрузили из хранилища. Прежде чем приступить к построению модели, проверьте по нашей инструкции их на корректность.

**Наш план:**

**1. Подготовим данные**

**2. Проанализируем данные**

**3. Построим модель**

**4. Опишем выводы.**

В этом проекте будут использованы следующие библиотеки:

pandas, seaborn, numpy, scikit-learn (sklearn).