# Постановка задачи

Примените знания к реальной задаче Data Science из золотодобывающей отрасли. Проект предоставлен компанией [«Цифра»](http://zyfra.com/).

Когда закончите, отправьте работу на проверку ревьюеру: он пришлёт комментарии в течение суток. После этого нужно доработать проект и пройти повторную проверку.

Скорее всего, вы будете дорабатывать кейс по комментариям ещё несколько раз. Это нормально.

Проект завершён, когда ревьюер одобрил все доработки.

# 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.

**Вам нужно:**

* Подготовить данные;
* Провести исследовательский анализ данных;
* Построить и обучить модель.

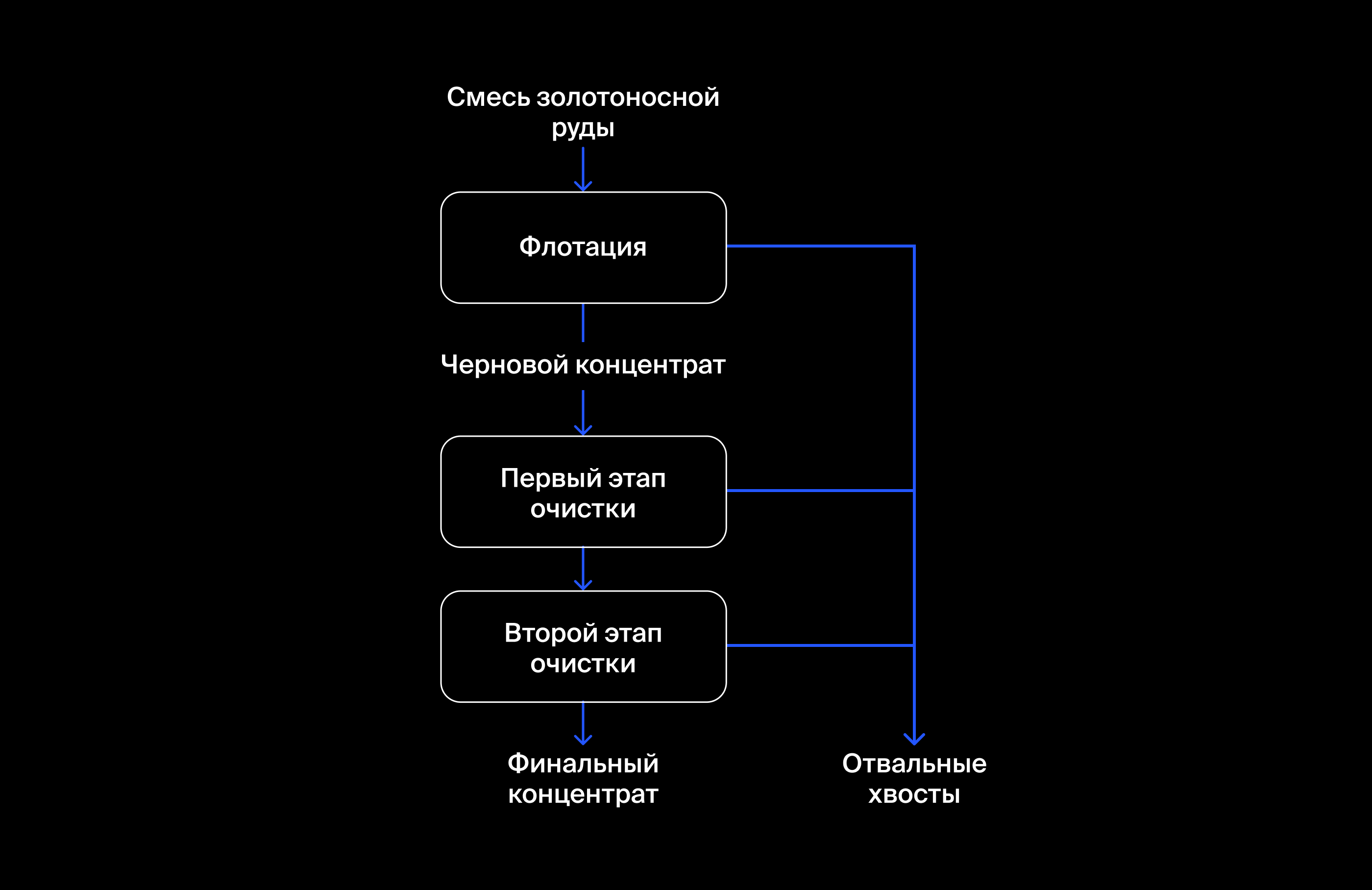
Чтобы выполнить проект, обращайтесь к библиотекам *pandas*, *matplotlib* и *sklearn.* Вам поможет их документация.

Следующий урок посвящён технологическому процессу очистки руды. Решите сами, какие детали нужны для построения модели, а какие — нет.

# Технологический процесс

Как золото получают из руды? Изучите внимательно стадии процесса.

Когда добытая руда проходит первичную обработку, получается дроблёная смесь. Её отправляют на флотацию (обогащение) и двухэтапную очистку.



Опишем каждую стадию:

**1. Флотация**

Во флотационную установку подаётся смесь золотосодержащей руды. После обогащения получается черновой концентрат и «отвальные хвосты», то есть остатки продукта с низкой концентрацией ценных металлов.

На стабильность этого процесса влияет непостоянное и неоптимальное физико-химическое состояние флотационной пульпы (смеси твёрдых частиц и жидкости).

**2. Очистка**

Черновой концентрат проходит две очистки. На выходе получается финальный концентрат и новые отвальные хвосты.

## 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

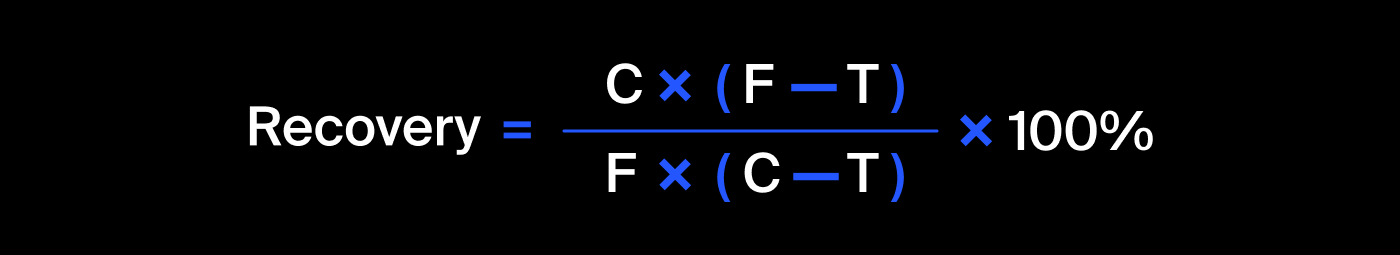
Возможные значения для блока [тип\_параметра]:

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

## Расчёт эффективности

Вам нужно смоделировать процесс восстановления золота из золотосодержащей руды.

Эффективность обогащения рассчитывается по формуле



где:

* *C* — доля золота в концентрате после флотации/очистки;
* *F* — доля золота в сырье/концентрате до флотации/очистки;
* *T* — доля золота в отвальных хвостах после флотации/очистки.

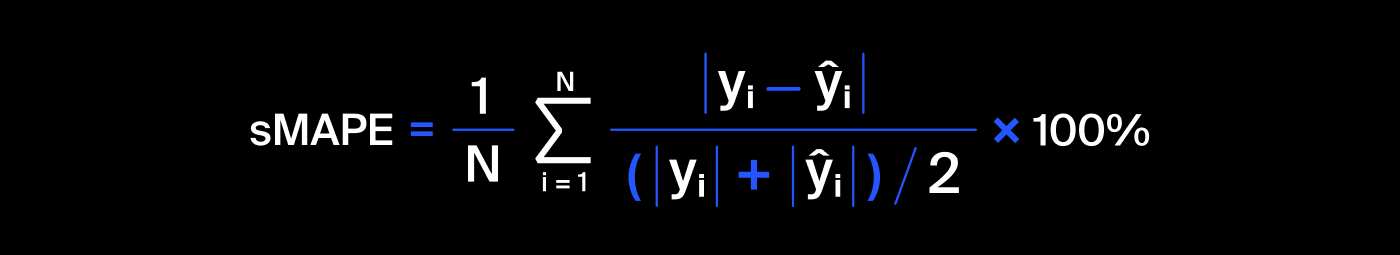
Для прогноза коэффициента нужно найти долю золота в концентратах и хвостах. Причём важен не только финальный продукт, но и черновой концентрат.

## Метрика качества

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

Она похожа на *MAE*, но выражается не в абсолютных величинах, а в относительных. Почему *симметричная*? Она одинаково учитывает масштаб и целевого признака, и предсказания.

Метрика *sMAPE* вычисляется так:



Обозначения:



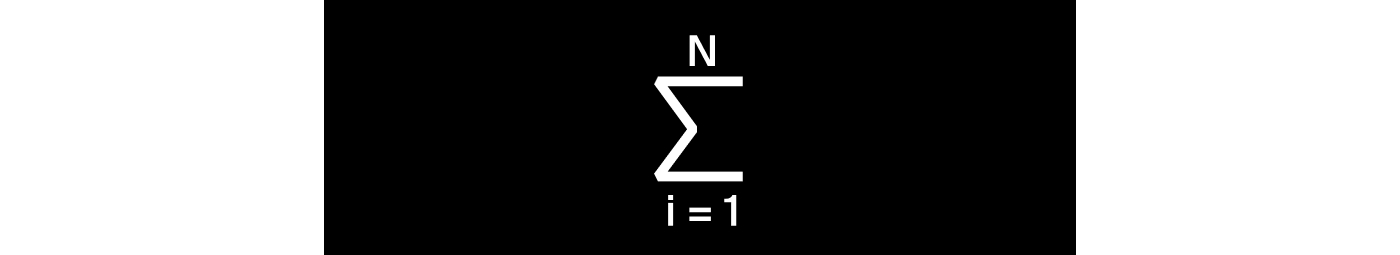
* Значение целевого признака для объекта с порядковым номером ***i*** в выборке, на которой измеряется качество.



* Значение предсказания для объекта с порядковым номером ***i***, например, в тестовой выборке.



* Количество объектов в выборке.

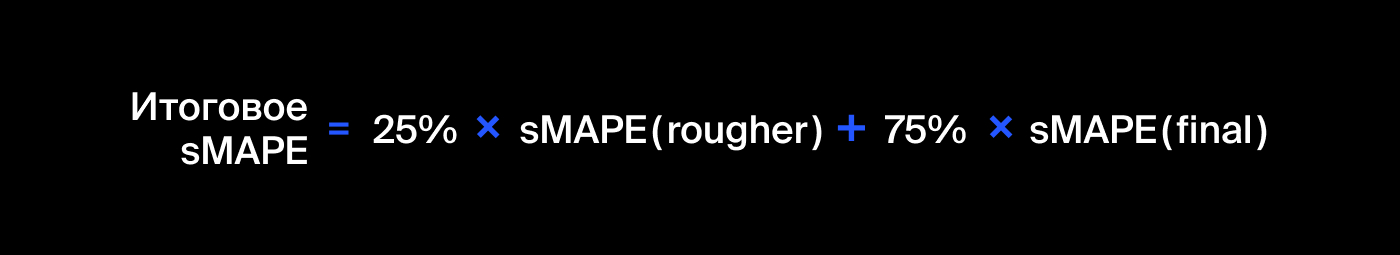


* Суммирование по всем объектам выборки (***i*** меняется от 1 до *N*).

Нужно спрогнозировать сразу две величины:

* эффективность обогащения чернового концентрата rougher.output.recovery;
* эффективность обогащения финального концентрата final.output.recovery.

Итоговая метрика складывается из двух величин:



# Описание проекта

Данные находятся в трёх файлах:

* gold\_recovery\_train\_new.csv — обучающая выборка;
* gold\_recovery\_test\_new.csv — тестовая выборка;
* gold\_recovery\_full\_new.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.

## Инструкция по выполнению проекта:

**1. Подготовьте данные**

1.1. Откройте файлы и изучите их.

Путь к файлам:

* */datasets/gold\_recovery\_train\_new.csv*.
* */datasets/gold\_recovery\_test\_new.csv*.
* */datasets/gold\_recovery\_full\_new.csv*.

1.2. Проверьте, что эффективность обогащения рассчитана правильно. Вычислите её на обучающей выборке для признака rougher.output.recovery.

Найдите *MAE* между вашими расчётами и значением признака. Опишите выводы.

1.3. Проанализируйте признаки, недоступные в тестовой выборке. Что это за параметры? К какому типу относятся?

1.4. Проведите предобработку данных.

**2. Проанализируйте данные**

2.1. Посмотрите, как меняется концентрация металлов (*Au, Ag, Pb*) на различных этапах очистки. Опишите выводы.

2.2. Сравните распределения размеров гранул сырья на обучающей и тестовой выборках. Если распределения сильно отличаются друг от друга, оценка модели будет неправильной.

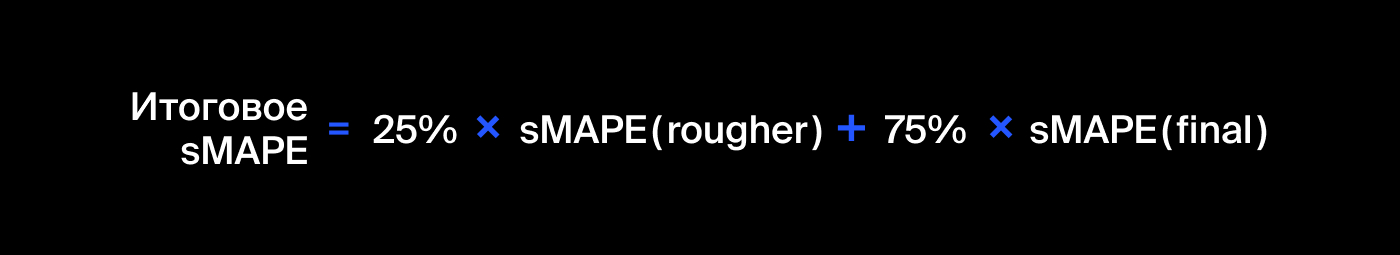
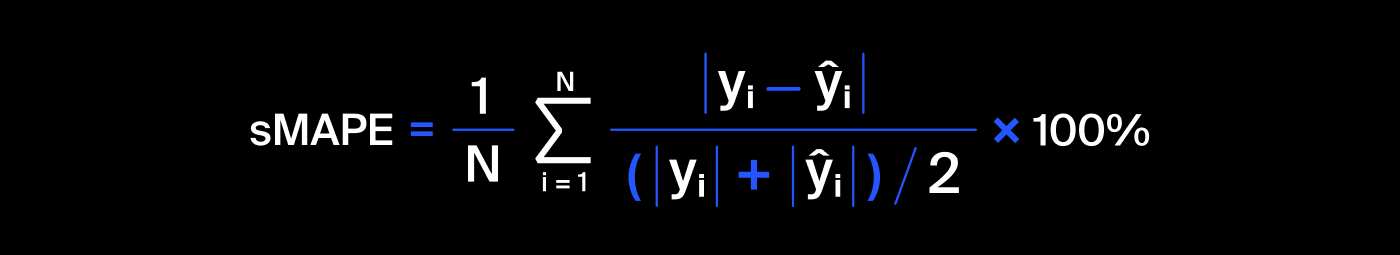
2.3. Исследуйте суммарную концентрацию всех веществ на разных стадиях: в сырье, в черновом и финальном концентратах.

**3. Постройте модель**

3.1. Напишите функцию для вычисления итоговой *sMAPE*.

3.2. Обучите разные модели и оцените их качество кросс-валидацией. Выберите лучшую модель и проверьте её на тестовой выборке. Опишите выводы.

Вам пригодятся формулы метрик качества:



**Как будут проверять проект?!**

Мы подготовили критерии оценки проекта, которыми руководствуются наставники. Прежде чем приступить к решению кейса, внимательно их изучите.

На что обращают внимание наставники, проверяя проект:

* Насколько хорошо вы готовите и анализируете данные?
* Какие модели строите?
* Как проверяете их качество?
* Все ли шаги по инструкции выполняете?
* Следите за структурой проекта и поясняете выполненные шаги?
* Какие выводы делаете?
* Поддерживаете аккуратность кода и не дублируете его?

Всё, что вам нужно знать, есть в шпаргалках и конспектах прошлых тем.

Успеха!