SIBUR Challenge 2020

Задача “Сырьё”

команда IV & Evteev
1 место public / 2 место private

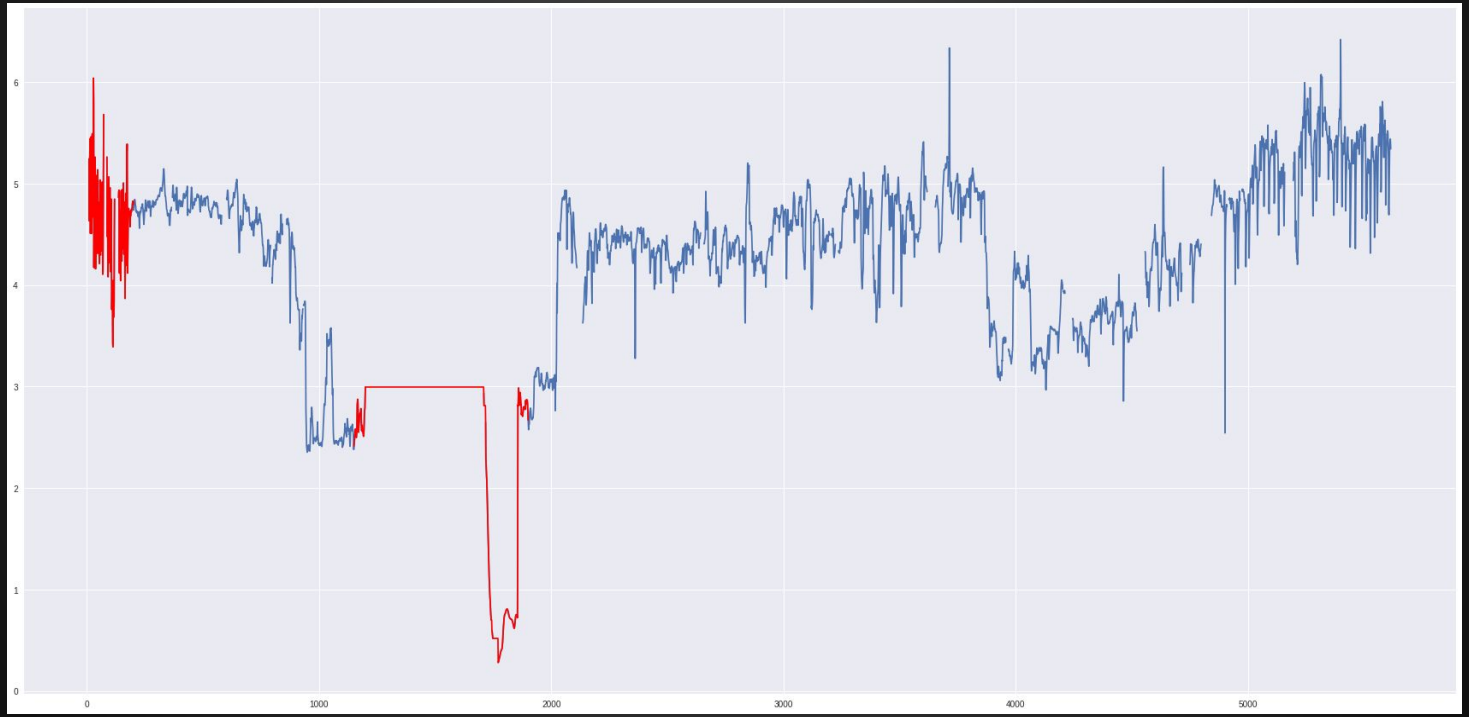
Игорь Кучеревский г. Москва
Владислав Евтеев г. Краснодар

Основные этапы решения

1. Очистка данных
    a. Грубая чистка
    b. Поиск выбросов
    c. Заполнение пропусков
    d. Корректировка процентных сумм элементов
    e. Визуальная оценка таргетов
2. Адаптивная система сдвигов
3. Обучение и постпроцессинг
    a. Оверсэмплирование
    b. Построение валидации
    c. Обучение
    d. Экспоненциальное сглаживание
4. Результаты

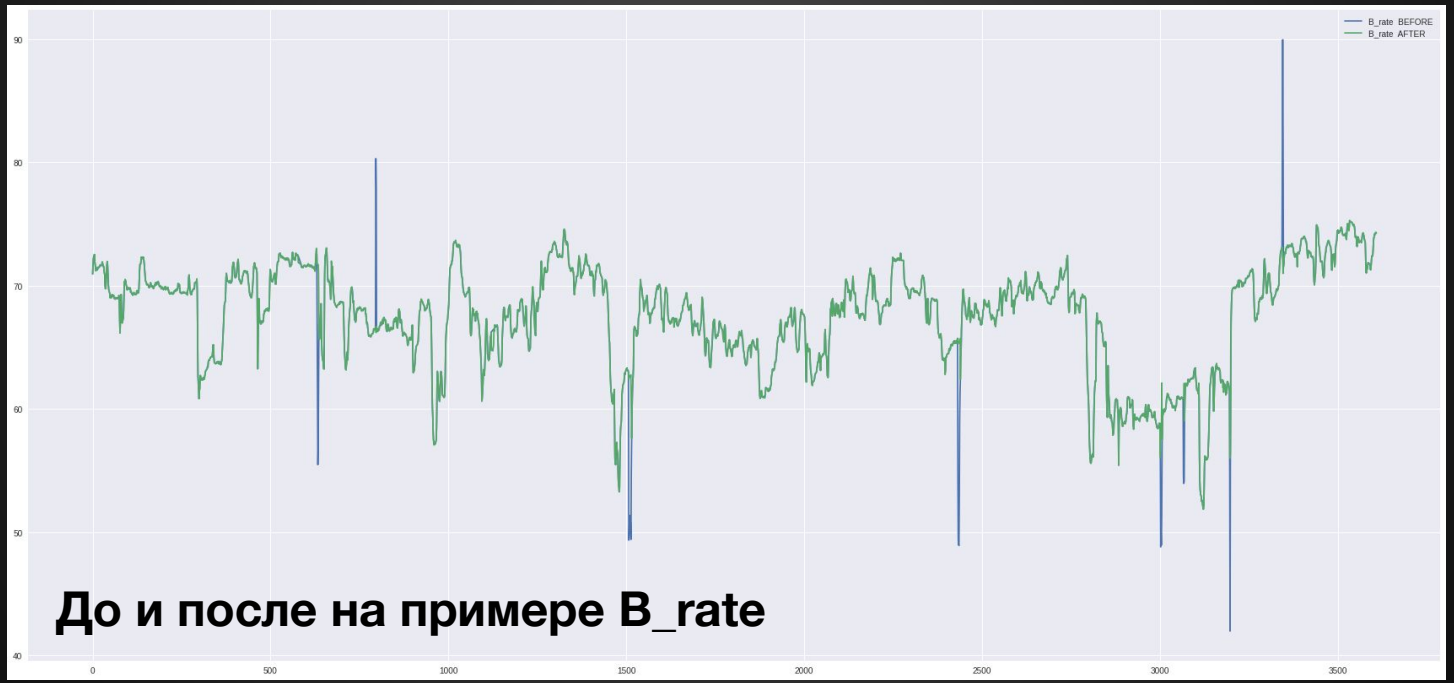
Очистка данных: грубая чистка

- **Из тренировочной выборки были сразу выброшены два интервала данных: это очень шумные первые 200 значений и огромный бассейн пропусков и выбросов в первой половине**



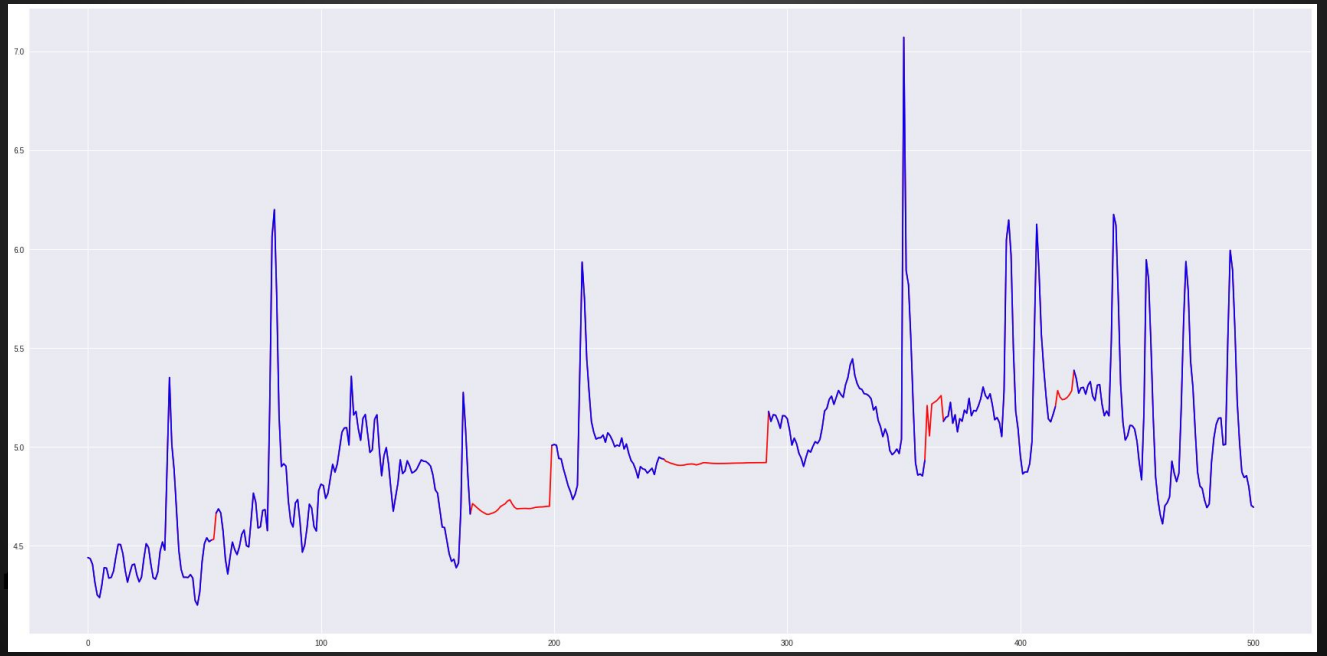
Очистка данных: поиск выбросов

- **В скользящем окне рассчитывалось среднее значение и среднеквадратичное отклонение. Если значение было больше mean+n*sigma или меньше mean-n*sigma, оно объявлялось как None.**



Очистка данных: заполнение пропусков

- При восстановление значений A_rate/B_rate, если значение по одному из признаков в паре было известно, строилась регрессия в скользящем окне и предсказывался второй
- Пропуски по хим. веществам заполнялись суммой среднего значения и среднего изменение в скользящем окне

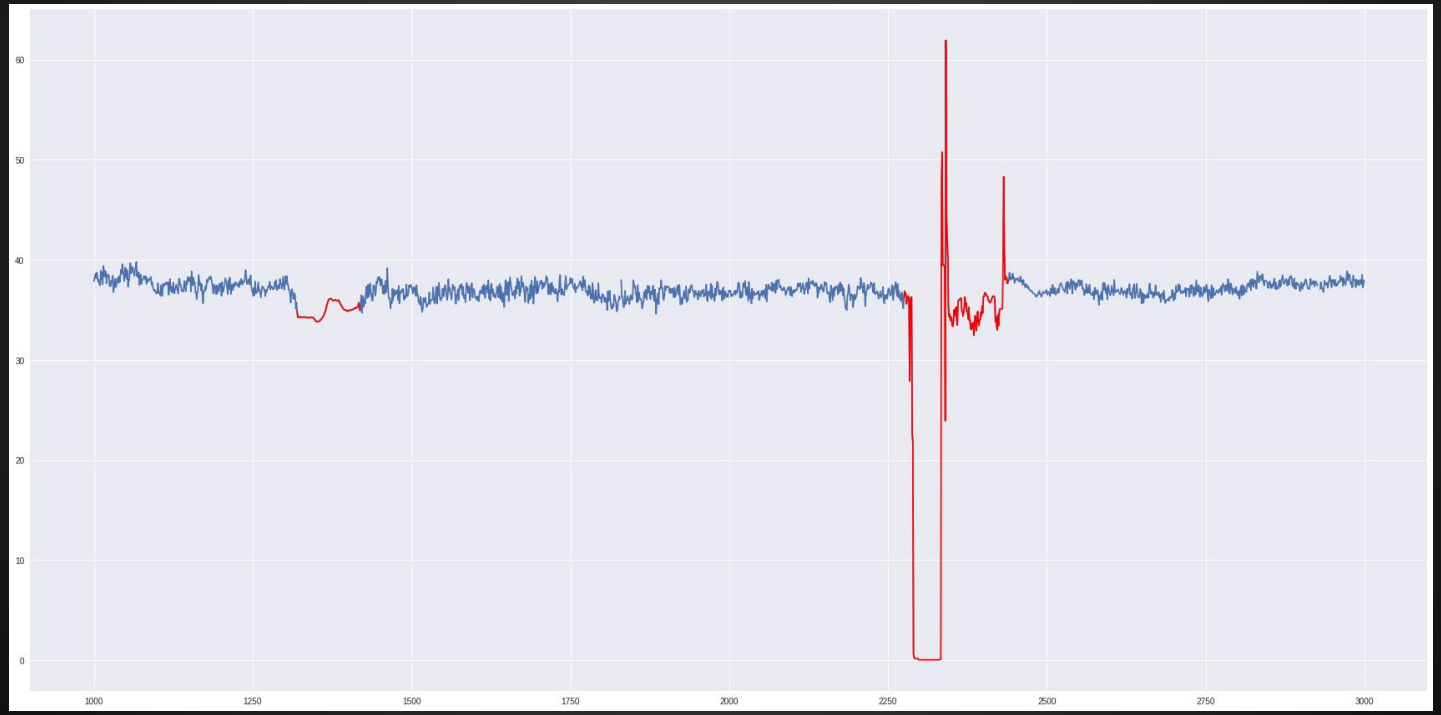


Очистка данных: корректировка процентов

- После восстановления пропусков, был проведен анализ суммы химических элементов. Если она выходила за логичные значения интервала [99.9, 100]. То пропорционально все значения элементов уменьшались или увеличивались, образуя сумму 99.95

Очистка данных: визуальная оценка таргетов

- После всех предыдущих шагов было проанализировано поведение таргетов и выброшено два интервала данных, странно сглаженный в первой половине и ужасно шумный во второй.



Адаптивная система сдвигов

- В конце соревнования была имплементирована адаптивная система сдвигов вместо константной.
- В скользящем окне размера 190 считался средний B_rate и по найденной на валидации пропорции высчитывался оптимальный сдвиг для каждого объекта. Как итог оптимальный сдвиг варьируется от 180 до 200 по всей выборке.

Константный сдвиг

Адаптивный сдвиг

Обучение и постпроцессинг: оверсэмплирование

- Так как данных для обучения после очистки осталось мало, было использовано оверсэмплирование
  1. Между двумя соседними значениями вставлялся пропуск
  2. Пропуски заполнялись интерполяцией

\	f1	f2
obj1	1	10
obj2	2	15
obj3	3	20
obj4	4	25

1

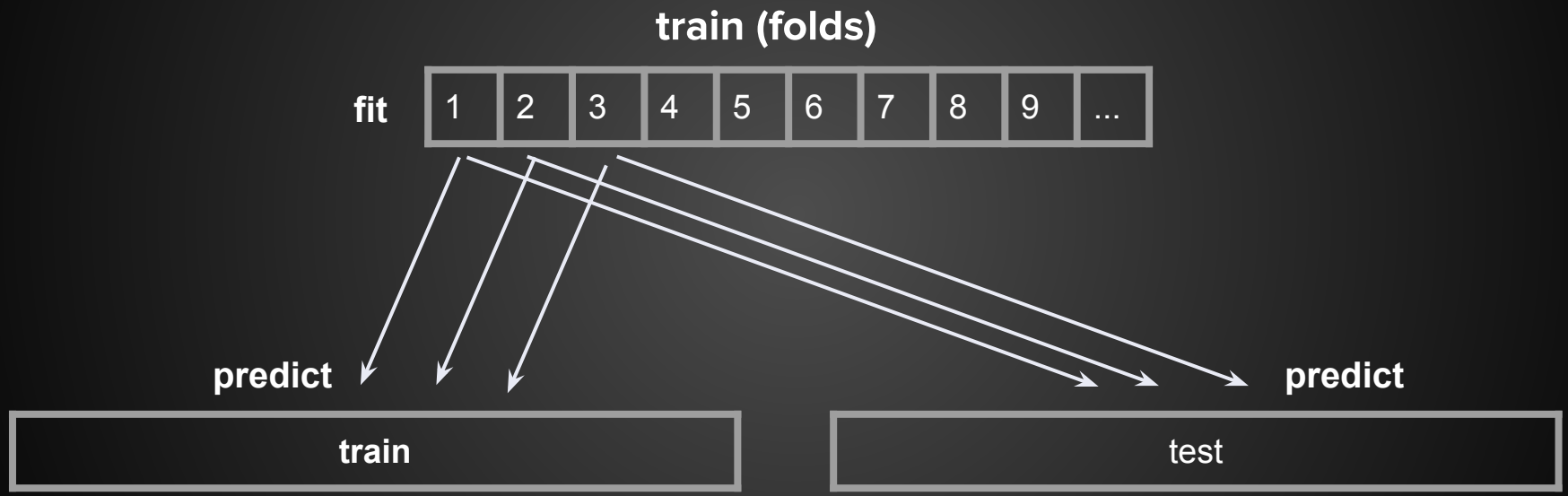
\	f1	f2
obj1	1	10
obj2	none	none
obj3	2	15
obj4	none	none
obj5	3	20
obj6	none	none
obj7	4	25

2

\	f1	f2
obj1	1	10
obj2	1.5	12.5
obj3	2	15
obj4	2.5	17.5
obj5	3	20
obj6	3.5	22.5
obj7	4	25

Обучение и постпроцессинг: валидация

- Была написана собственная система кросс-валидации. Тренировочная выборка делилась на 16 пересекающихся отрезков. Модель обучалась на каждом из них и предсказывала весь train и весь test. Затем полученные значения усреднялись.



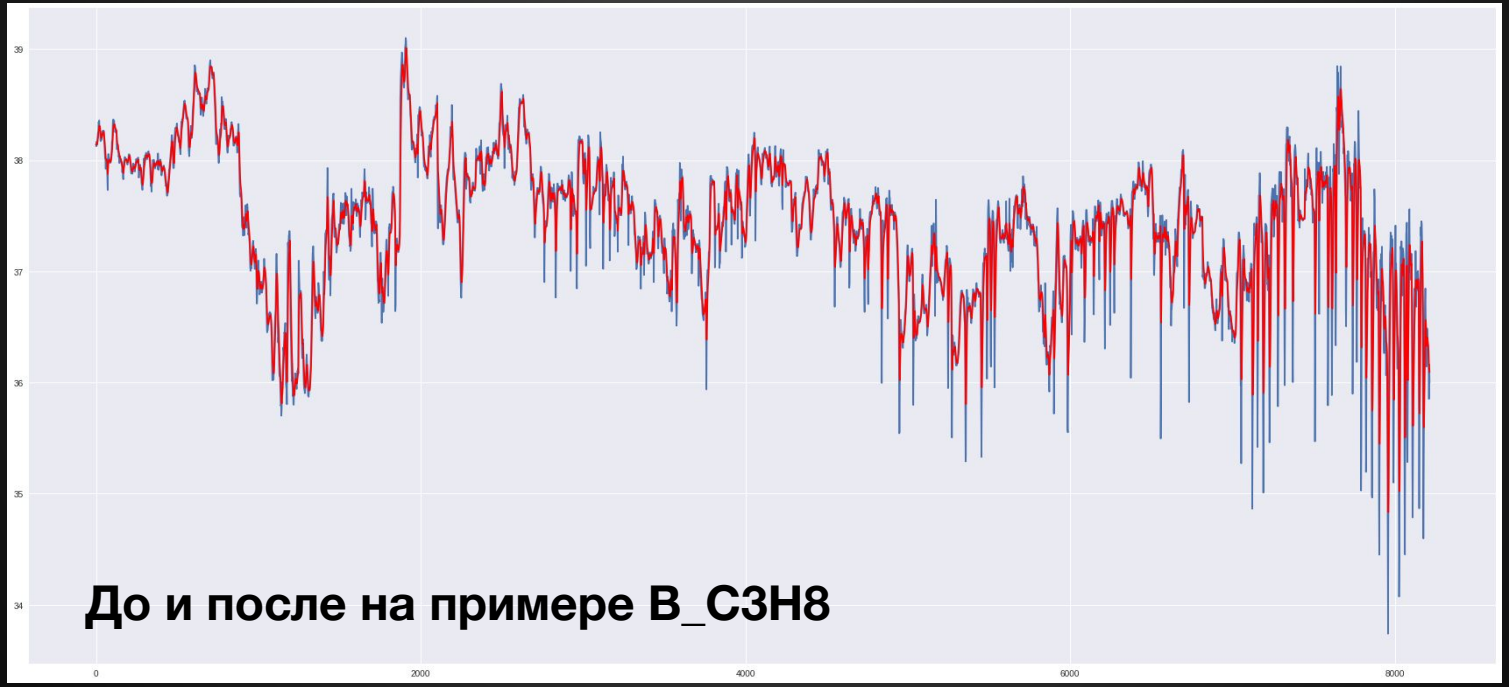
Обучение и постпроцессинг: обучение

- Были применены градиентные бустинги, опорные вектора на разных ядрах, перебран весь sklearn, ансамблирование и стэкинг, но результата это не дало
- Поэтому в итоговом решение использовалась линейная регрессия с L2 регуляризацией на базовых 10 признаках

```
model = Ridge()
model.fit(train, target)
result = model.predict(test)
```

Обучение и постпроцессинг: экспоненциальное сглаживание

- После получения предсказаний, чтобы уменьшить влияние шумов и выбросов, к каждому таргету было применено экспоненциальное сглаживание.



Приблизительные результаты этапов

параметры модели	public result	private result
мл на сырых данных	3.7281	4.2757
чистка и константный сдвиг на 195 только трейна	2.9316	3.6011
чистка и константный сдвиг на 195 всех данных	2.0913	3.0464
добавление фолдов и константный сдвиг на 192	1.7657	2.9175
добавление оверсэмплинга	1.7042	2.7880
добавление адаптивных сдвигов	1.4586	2.6632
добавление экспоненциального сглаживание	1.4205	2.6509

Что дало хороший прирост ?

- Подбор системы кросс-валидацит
- Исключение аномальных участков обучающей выборки
- Применение адаптивных сдвигов
- Постпроцессинг, экспоненциальное сглаживание предсказаний