

Создание классификатора фармакометрических профилей для выявления несогласованностей между дозой и концентрацией

Федотов Влад, MSc-2 LS, Сколтех

Мотивация

Чистота данных фармакометрики играет большую роль для анализа, так как от неё напрямую зависит качество подобранных параметров для модели.

В силу этого важной задачей является контроль качества приходящих данных. Проблема такого рода называется **задачей детекции аномалий**.

Подходы к решению

Для данной задачи можно выделить два основных пути. Условия для них отличаются только в одном, а именно, в том, знает ли модель при обучении какие данные относятся к ошибкам, а какие являются правильными:

- Supervised метод (обучение на заранее размеченных классах, с учителем)
- Unsupervised метод (когда классы заранее не показываются модели, без учителя)

Плюсом supervised моделей является, как правило, их высокая точность, так как обучение ведется напрямую через демонстрацию ошибочных данных. В то же время unsupervised модели приятны тем, что не требуют разметки, то есть могут обучаться на сырых данных.

Подходы к решению

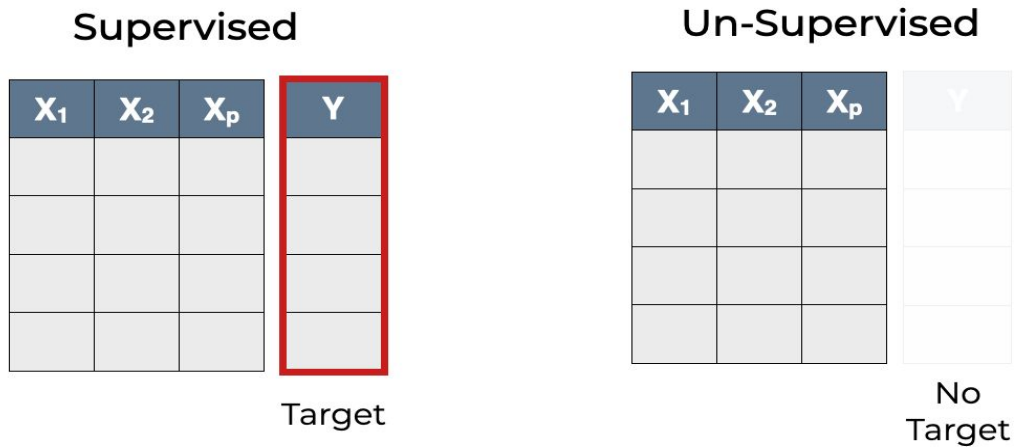


Рис. 1: Схема данных в различных алгоритмах машинного обучения

Генерация данных

Данные для работы были сгенерированы с помощью библиотеки RxODE, применяемой в языке R.

В качестве датасета, на котором будут проверяться модели, был сделан набор профилей, состоящих из шести точек, представляющих собой замеры концентрации в момент подачи пациенту дозы (C_{min}). Также у каждой точки было отмечено подавалась ли доза в данный момент.

Всего для анализа было создано 2240 профилей пациентов, из которых 240 содержат ошибки.

Для симуляции индивидуальной вариации были использованы параметры, подобранные в экспериментах.

Подробнее о модели

$$\begin{cases} \frac{dC_{depot}}{dt} = -k_a C_{depot} & : \text{Компартмент дозы} \\ \frac{dC_{center}}{dt} = k_a C_{depot} - CL \frac{C_{center}}{V_1} - Q \left(\frac{C_{center}}{V_1} - \frac{C_{perif}}{V_2} \right) & : \text{Центральный компартмент} \\ \frac{dC_{perif}}{dt} = Q \left(\frac{C_{center}}{V_1} - \frac{C_{perif}}{V_2} \right) & : \text{Периферический компартмент} \end{cases}$$

$$\begin{cases} k_a = 0.22 & : \text{Константа абсорбции} \\ \langle CL \rangle = 0.18 & : \text{Клиренс} \\ \langle V_1 \rangle = 4.1 & : \text{Объём центрального компартмента} \\ Q = 0.59 & : \text{Межкомпартментный клиренс} \\ V_2 = 2.33 & : \text{Объём периферического компартмента} \end{cases}$$

Рис.2: Структурная модель и параметры системы

Подробнее о модели

$\omega_{V_1} = 0.42$: Дисперсия межиндивидуальной вариабельности объёма распределения
$\omega_{CL} = 0.73$: Дисперсия межиндивидуальной вариабельности клиренса
$V_{1i} = \langle V_1 \rangle \exp(\omega_{V_1})$: Объём с учётом вариабельности
$CL_i = \langle CL \rangle \exp(\omega_{CL})$: Клиренс с учётом вариабельности

Рис.3: Статистическая модель для параметров

Примеры сгенерированных профилей

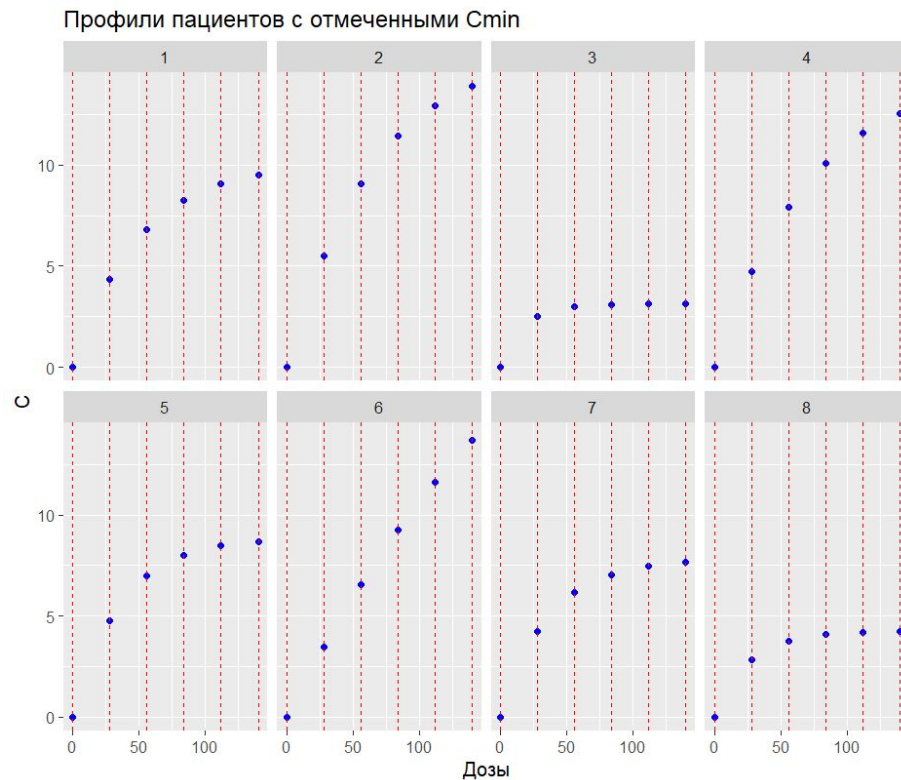


Рис.4: Симулированные профили

Подробнее об аномалиях

- **Тип 1** — при симуляции РК профилей одна из доз была пропущена, однако в дозировках переданного датасета отражено этого не было;
- **Тип 2** — РК профили генерировались с регулярно расположенными равными дозами, однако в дозировках переданного датасета указано, будто одна из доз пропущена;
- **Тип 3** — РК профили генерировались так, что одна из доз в n раз больше, чем остальные, однако соответствующее dose escalation не зафиксировано в датасете — в колонке AMT.

Подробнее об аномалиях

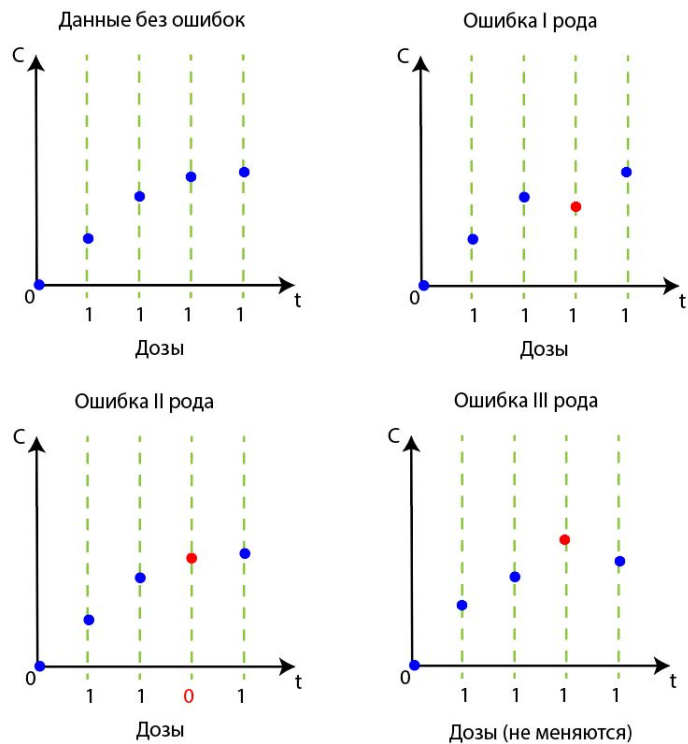


Рис.5: Типы классифицируемых аномалий

Метрики

В качестве метрики в силу большой несбалансированности классов был выбран F1 показатель, который является средним гармоническим от precision и recall.

Эти две метрики также дают важную информацию о работе модели, так как по сути показывают какой процент от угаданных моделью профилей составляют истинно плохие и какой процент от всех истинно плохих классов модель угадала соответственно.

Баланс классов

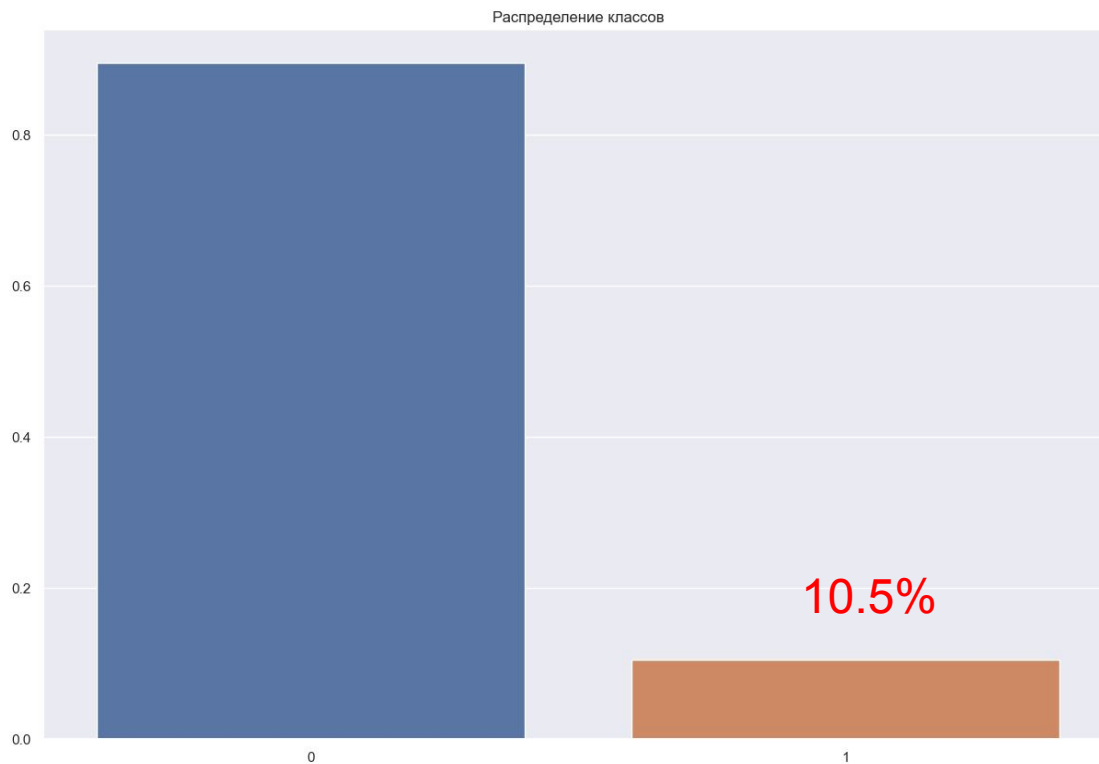


Рис.6: Распределение классов в данных

Метрики

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

GSDN 2011-2012

Рис.7: Метрики, используемые для оценки моделей

Supervised модели

В качестве моделей, применяемых с размеченными заранее классами, были выбраны классические ML модели:

- Naive Bayes
- Decision tree
- Random forest
- Gradient boosting classifier
- Logistic regression

Все эти модели достаточно известны и широко применяются для решения задач классификации.

Настройка гиперпараметров

Для улучшения работы моделей для каждой из них был проведен поиск по сетке гиперпараметров, чтобы подобрать наилучшую их комбинацию. В результате проведения подобной процедуры, удалось увеличить показатели в более чем полтора раза.

В дальнейшем лучшие варианты моделей были протестированы на датасете, используя технику кросс-валидации для объективной оценки качества предсказаний.

Настройка гиперпараметров

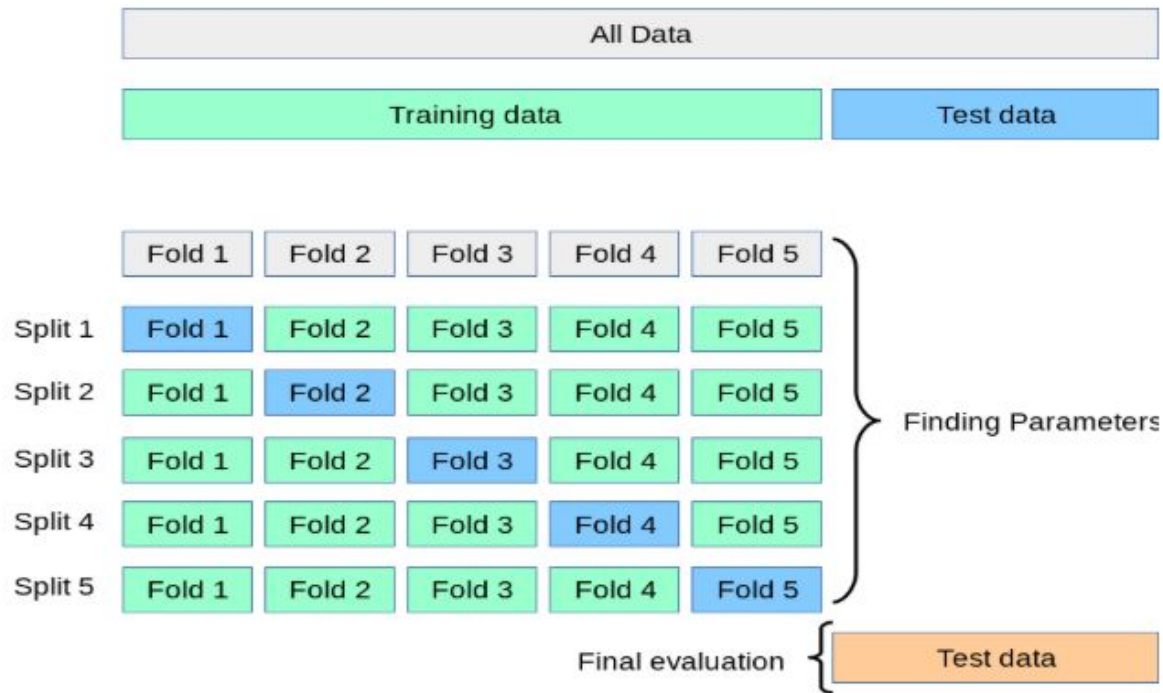


Рис.8: Схема кросс валидации с подбором параметров

Результаты кросс-валидации

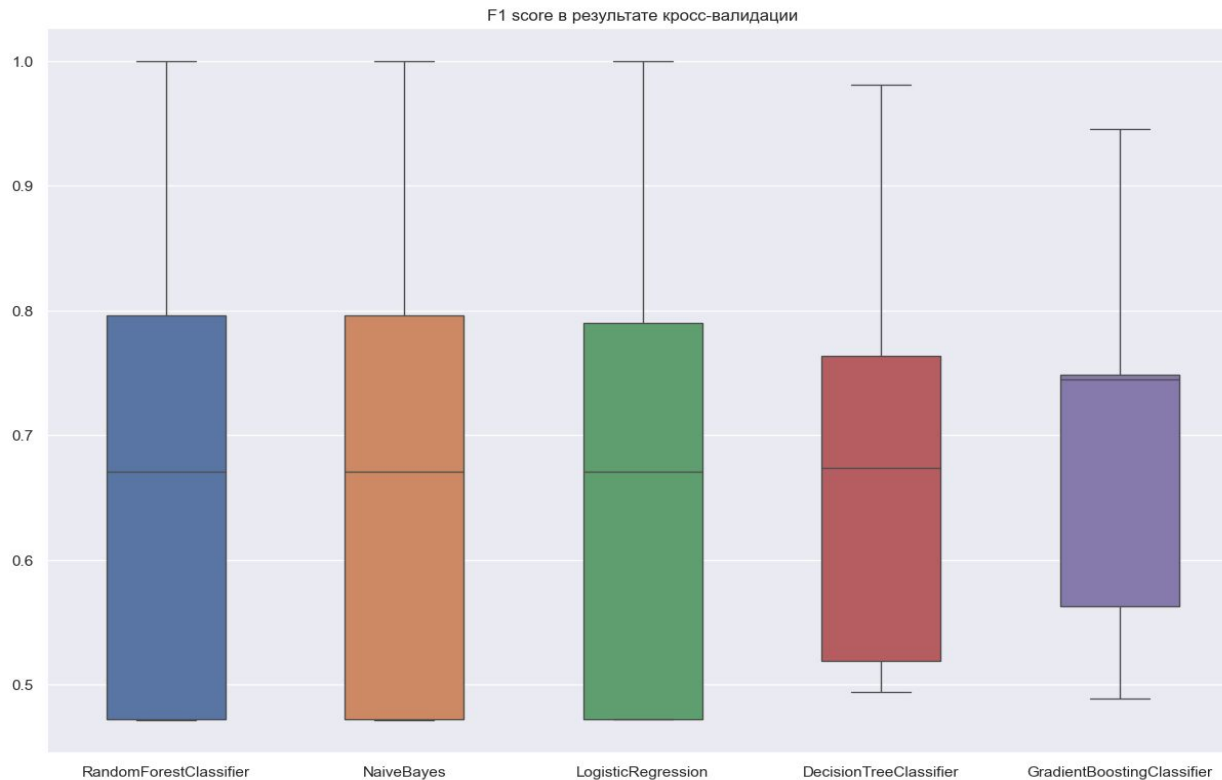


Рис.9: Результаты кросс-валидации для ML моделей

Кривые ROC

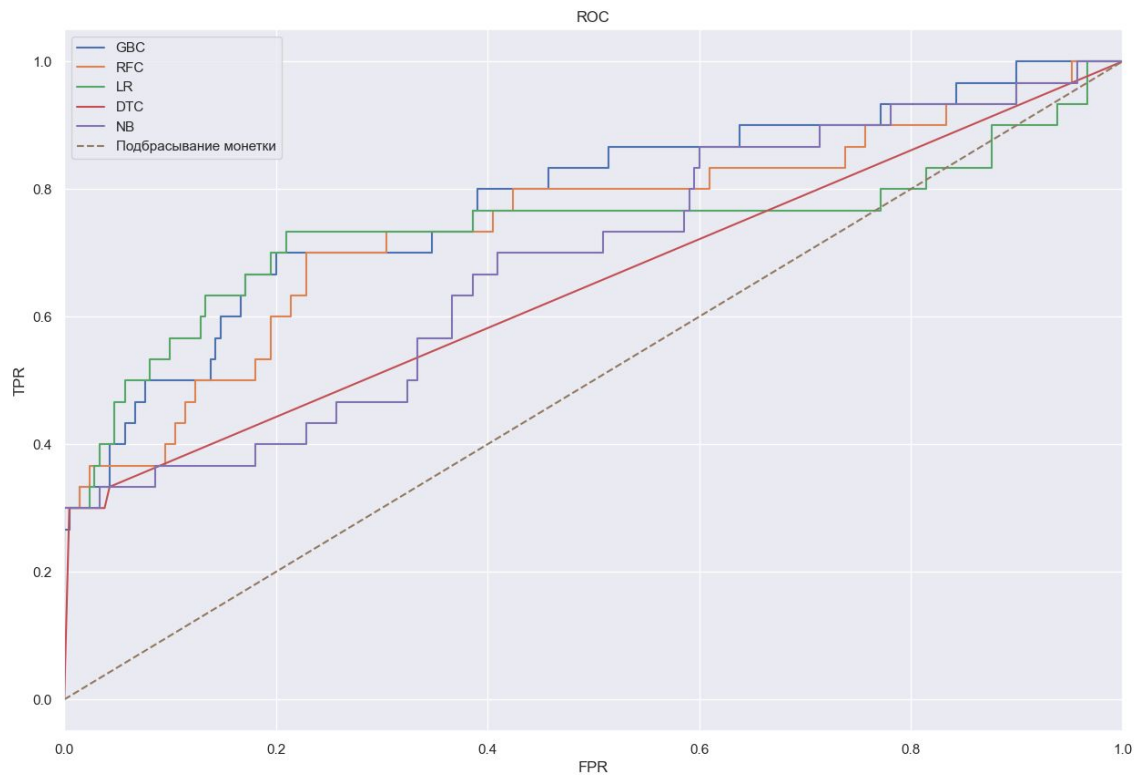


Рис.10: ROC кривые моделей

Unsupervised подход

В силу того, что данные поступают без указаний на то, какие из них ошибочны, более рабочим вариантом для реальной ситуации является unsupervised модель.

Она не требует разметки классов и обучается по совершенно другому принципу. Если supervised модель учится угадывать какие профили являются ошибочными, пользуясь заранее представленными шаблонами и участь по ним, то unsupervised использует подход генеративных моделей и старается воспроизвести хорошие профили максимально точно.

Дальше, после обучения, остаётся только выбрать критерий, по которому можно будет отличить хорошие профили от плохих.

Unsupervised подход

В качестве архитектуры для unsupervised метода был выбран вариационный автокодировщик. Данная модель имеет достаточно широкое применение в генерации изображений или других данных.

Для того, чтобы модель правильно понимала наши данные, которые являются временными рядами, в качестве основных блоков были выбраны рекуррентные нейронные сети, так как они показывают лучший результат в работе с временными рядами, не считая трансформеров.

Архитектура предложенной модели

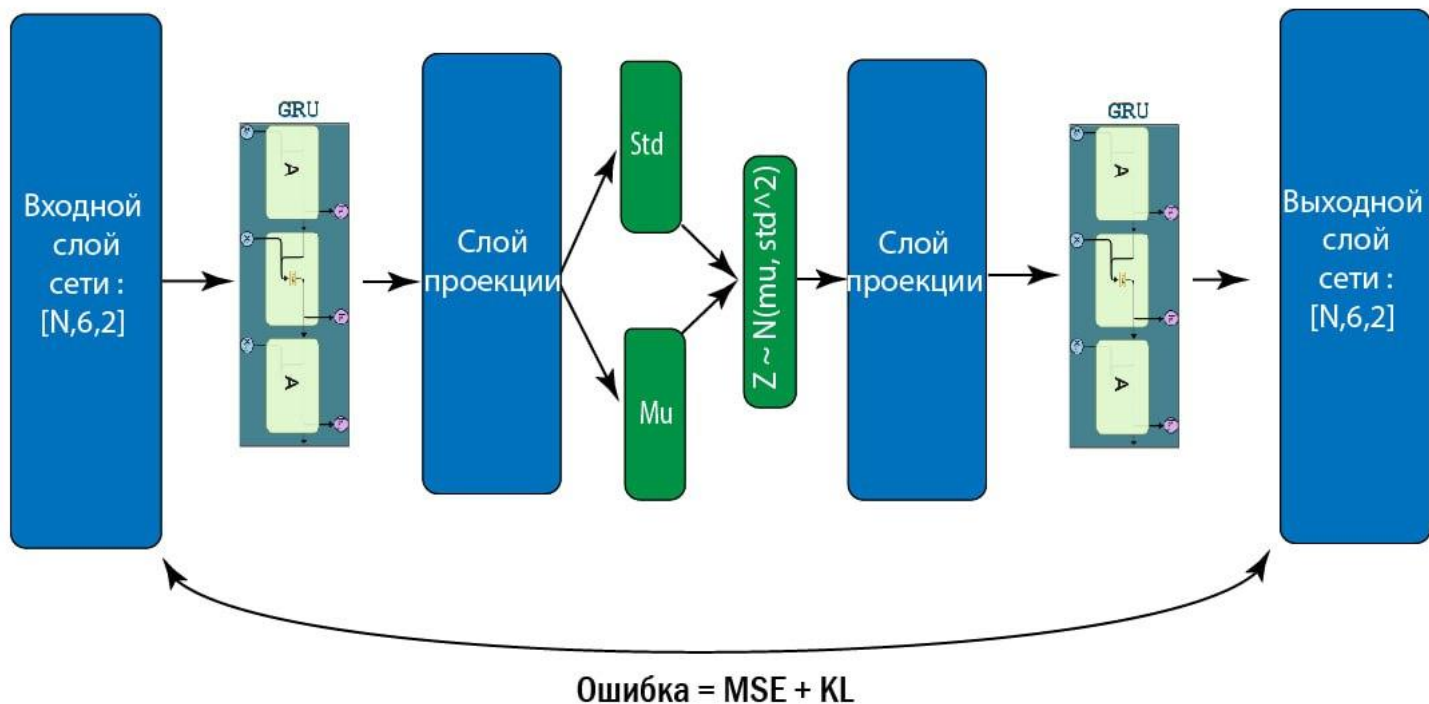


Рис.11: Архитектура RNN-VAE

Обучение VAE

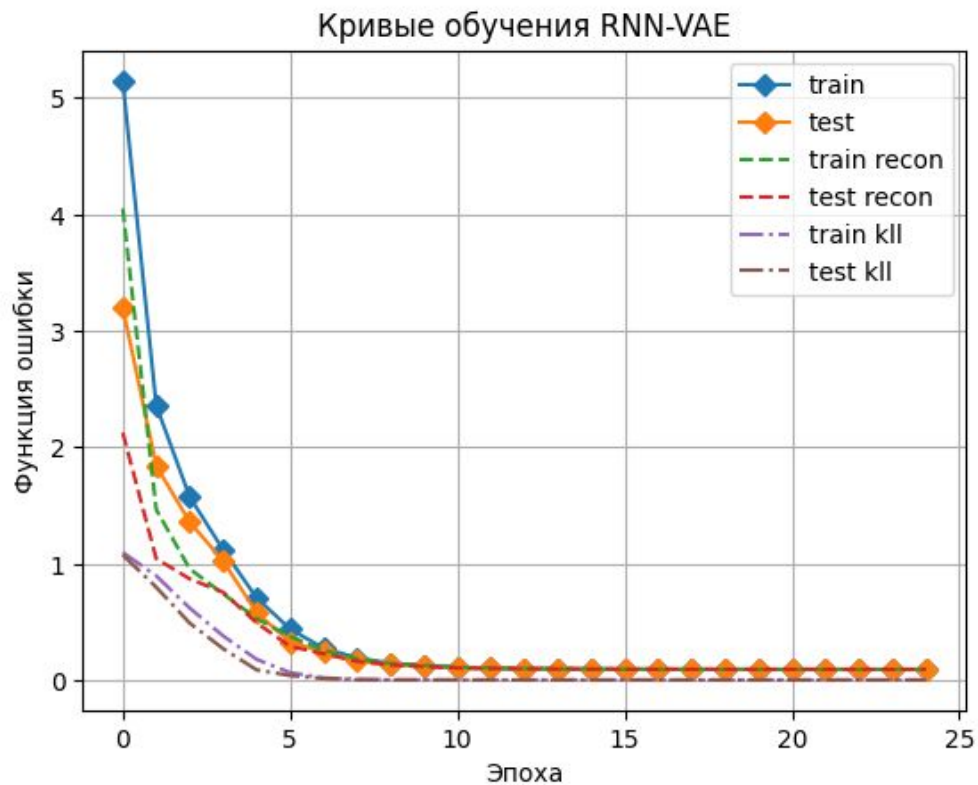


Рис.12: Кривые обучения для VAE

Unsupervised классификация

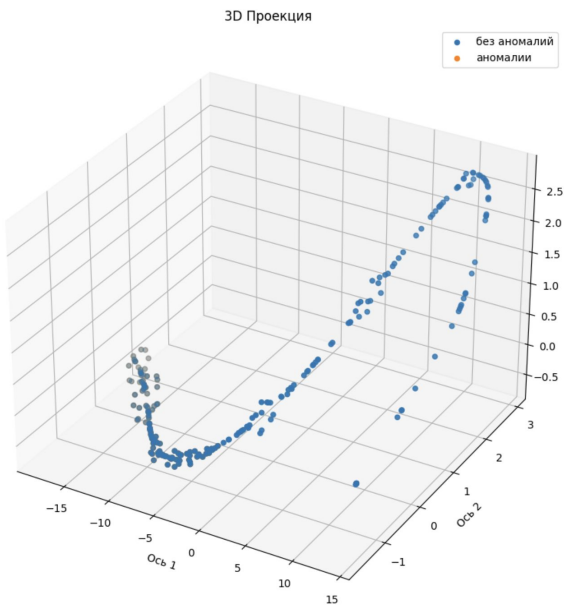
Для того, чтобы определить какие профили из реконструированных являются ошибочными, было решено исследовать пространство скрытых переменных.

Из него и из пространства ошибок реконструкции с помощью декартова произведения составлялось новое множество точек, которое имеет более высокую способность к кластеризации.

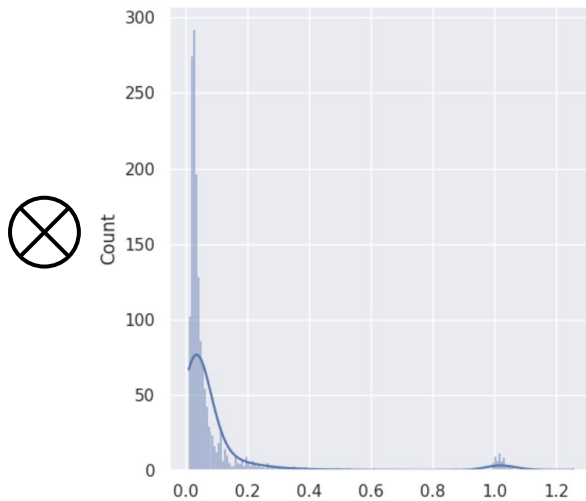
Дальше применялась кластеризация алгоритмом KMeans для разграничения групп профилей.

Скрытое представление для классификации

Скрытое пространство



Пространство ошибок



Результирующее пространство

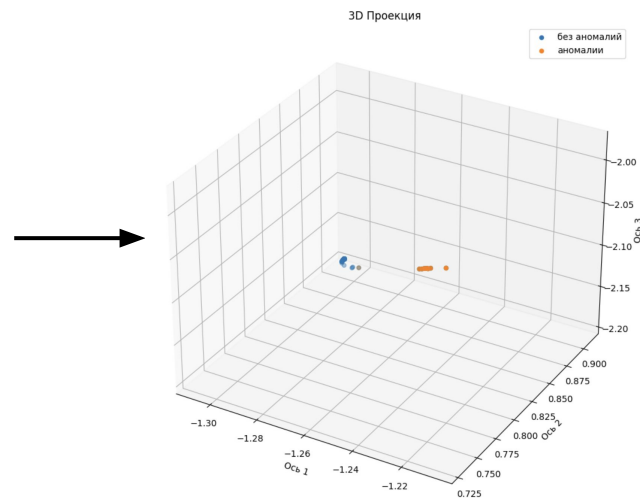


Рис.13: Схема подготовки скрытого пространства для классификации

Пространство скрытых переменных. PCA

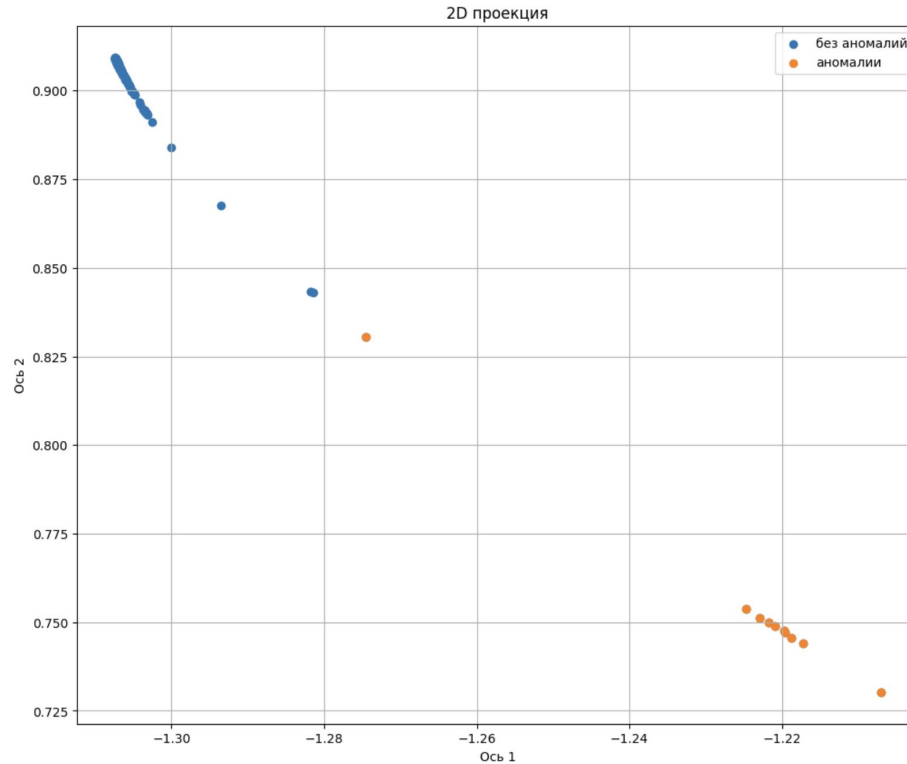


Рис.14: PCA проекция скрытого пространства

Пространство скрытых переменных.РСА

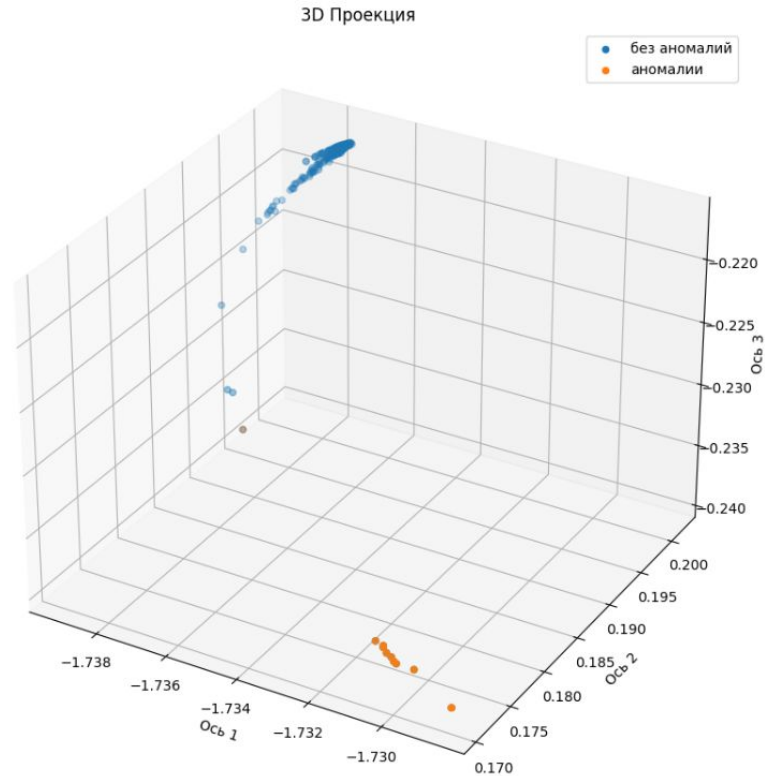


Рис.15: Трёхмерная РСА проекция

Пространство скрытых переменных. TSNE

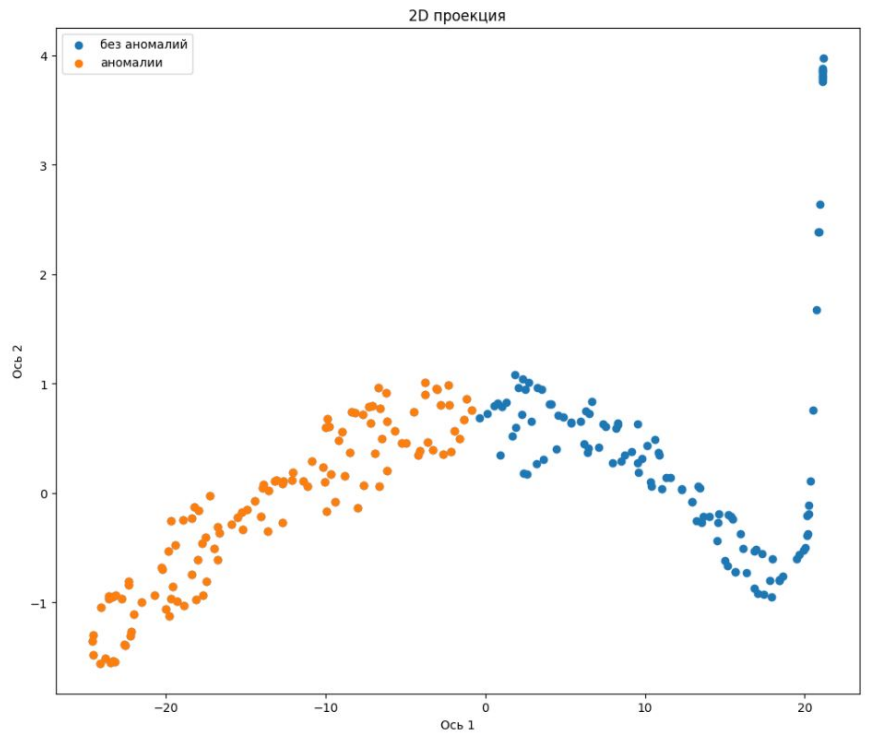


Рис.16: TSNE проекция скрытого пространства

Нахождение аномальных измерений

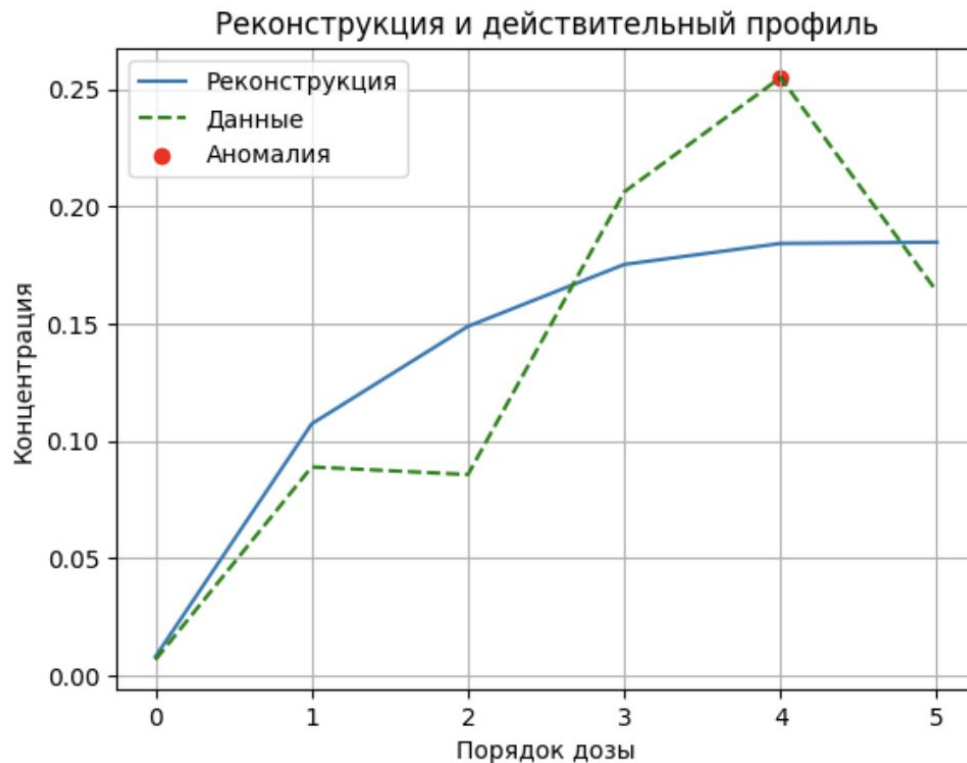


Рис.17: Определение аномальных измерений

Результаты

Среди классических ML моделей, используя supervised подход, наилучшим образом себя показал Gradient Boosting со средним **F1 близким к 0.7**.

В то же время unsupervised подход с использованием вариационного автокодировщика дает **средний F1 = 0.73**. Данный результат, хоть и не намного, но превосходит классический supervised подход.

Результаты. Среднее F1

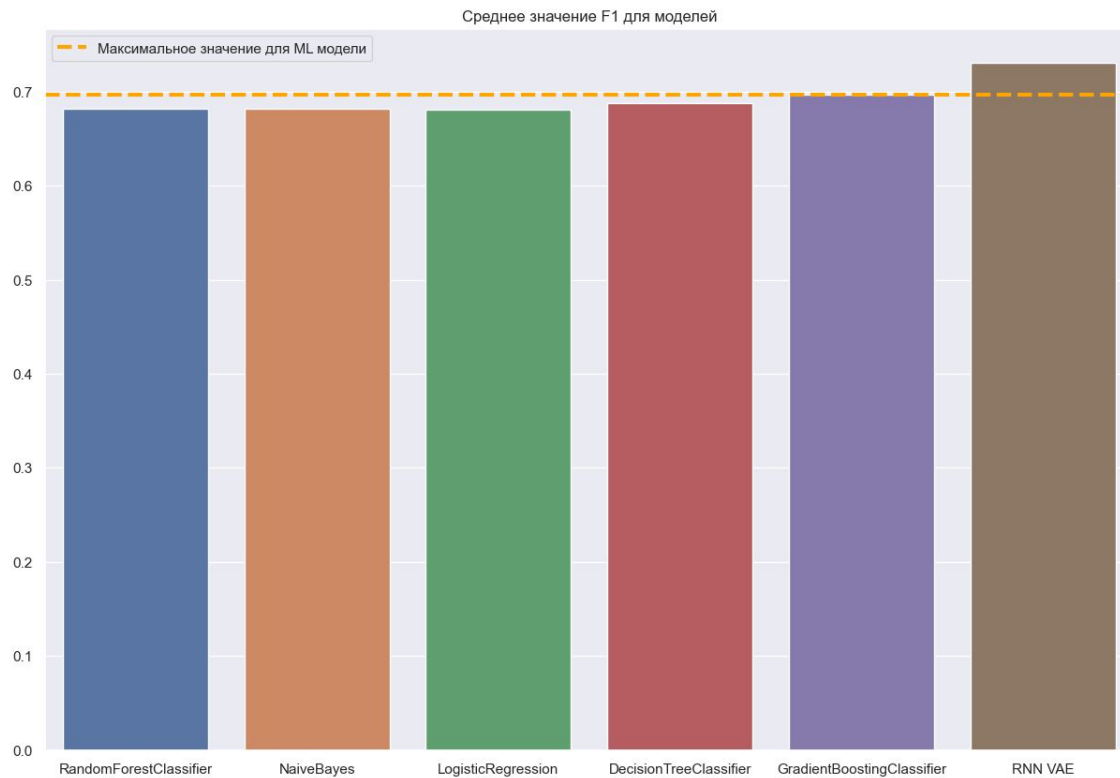


Рис.18: Среднее F1 для моделей

Значение ROC-AUC

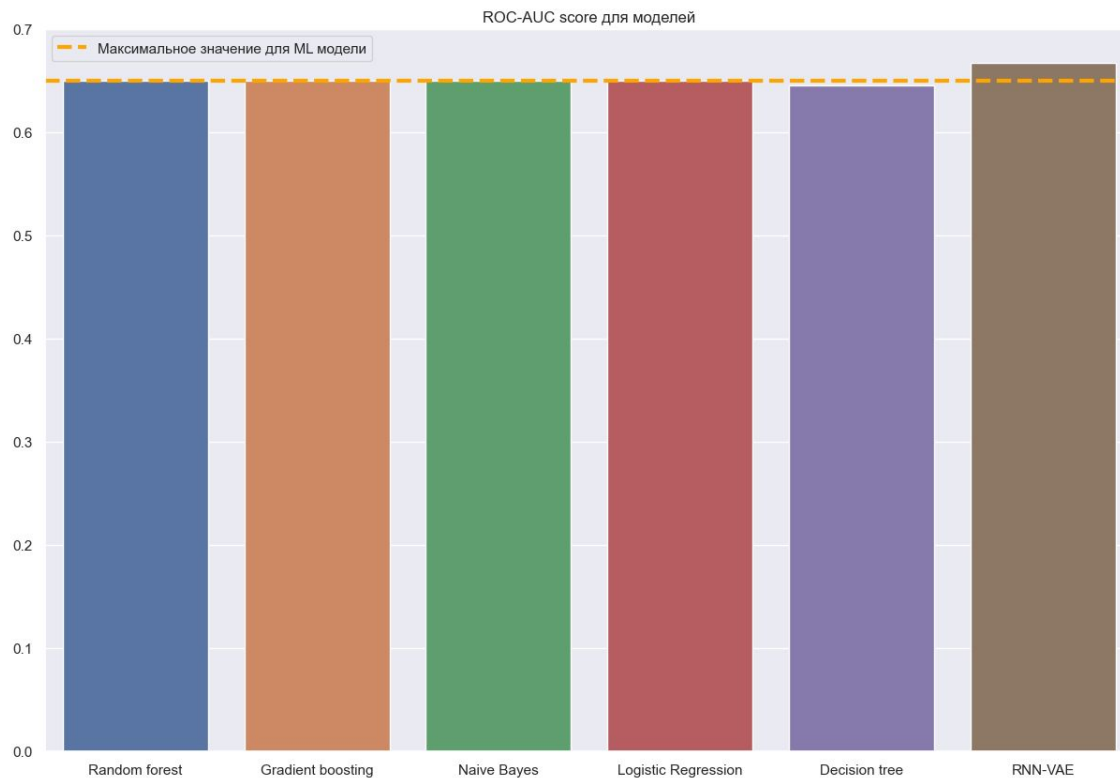


Рис.19: ROC-AUC моделей

Результаты

В то же время довольно важным замечанием будет то, что предложенная модель хуже справляется с классификацией в датасете, в котором представлен только третий тип ошибок.

F1 в данном случае равен 0.51, в отличие от supervised моделей, в которых этот показатель **может достигать до 0.57**.

Зато исключение из датасета этого типа ошибки приводит опять же к **precision = 1.0** и **F1 = 0.8**.

Заключение

В результате выполнения данной работы были предложены два варианта решения задачи детекции аномалий для фармакометрических данных: supervised метод, использующий классические алгоритмы ML, и unsupervised метод, для которого была создана нейронная сеть.

Заключение

В качестве улучшений для текущего алгоритма можно рассмотреть следующие шаги:

- Расширить семплинг (включать промежуточные точки, а не только S_{\min})
- Сделать алгоритм не зависящим от режима дозирования
- Протестировать модель на произвольных структурных моделях
- Улучшить работу модели, используя либо другие архитектуры, либо методы регуляризации KL функции ошибки, так как VAE часто подвергается “коллапсу” апостериорного распределения

Спасибо за внимание!

Связь VAE и EM-алгоритма

