**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**
**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Центр непрерывного образования
Факультета компьютерных наук

# ИТОГОВЫЙ ПРОЕКТ

Обнаружение аномалий в данных телеметрии УЭЦН

Выполнил (а):

Острик П.В.

Ф.И.О.

Руководитель:
Абдуракипов С.С.

Ф.И.О.

Москва
2021

Оглавление

## Раздел I. Введение

Сегодня более 80% процентов нефти добывается при помощи центробежных насосов (**УЭЦН** – установка электроприводного центробежного насоса). Многие из них в дополнение оснащены термоманометрическими системами, которые позволяет в реальном времени контролировать давление на сопле насоса и температуру двигателя насоса. Развитие технологий обработки данных позволило собрать ценную информацию о технологическом процессе. Ставшие широко распространенными технологии машинного обучения сделали возможным развитие эффективных предиктивных моделей на основе исторических данных. Как пример – сбор и обработка данных термоманометрических систем используя продвинутые алгоритмы анализа данных и машинного обучения позволяет получить важную информацию о процессе гидродинамического тестирования скважин, чтобы производить диагностирование насосов во время работы.

Еще один пример, который рассматривается в данной работе – снижение количества отказов оборудования. Для того, чтобы этого достичь, необходимо предотвращать работу оборудования в режимах с повышенной вероятностью отказа. Для этого необходимо проанализировать показатели датчиков на сопле насоса, находящегося в скважине. Однако проблема осложняется тем, что не все насосы оснащены подобными датчиками. Поэтому в данной работе мы будем рассматривать данные с показателей той части насосного оборудования, что расположена на поверхности. Данные показатели, в отличие, от данных с погруженного сопла, доступны для снятия на всех электрических центробежных насосах. Научившись выделять аномалии в показателях, мы сможем составить реестр потенциальных нештатных режимов работы оборудования, которые после валидации техническим специалистом могут быть положены в основу системы с обратной связью для предотвращения выхода оборудования из строя.

В рамках данной работы будут рассмотрены различные модели поиска аномалий (обучения без учителя), и проведен их сравнительный анализ, причем

в отличие от широко распространённого способа поиска аномалий по одному параметру (одномерный анализ), мы будем рассматривать множество параметров одновременно (многомерный анализ).

## Раздел II. Входные данные. Предобработка

Входные данные представляет собой матрицу размером 160262x18 – 160 тысяч наблюдений с 16 параметрами. Данные распределены по 17 насосным установкам со средним числом наблюдений в 9427 на каждый насос. Индексом является временная метка (Timestamp) вида гггг-мм-дд чч:мм:сс.

Параметры датасета представлены в таблице ниже.

Таблица №1. Параметры датасета

| Название параметра | Описание |
|---|---|
| engine | Наработка двигателя с момента последнего включения, сек |
| engine_idle | Время простоя двигателя с момента последнего выключения |
| workload_ch_h | Средняя скорость изменения загрузки двигателя ЧАС, %/час |
| workload_ch_d | Средняя скорость изменения загрузки двигателя СУТ, %/час |
| workload_ch_w | Средняя скорость изменения загрузки двигателя НЕДЕЛЯ |
| curr_ch_h | Средняя скорость изменения тока фазы А двигателя в ЧАС, А/час |
| curr_ch_d | Средняя скорость изменения тока фазы А двигателя в СУТ, А/час |

| | |
|---|---|
| curr_ch_w | Средняя скорость изменения тока фазы А двигателя в НЕДЕЛЯ, А/час |
| press_ch_h | Средняя скорость изменения давления в коллекторе ИУ в ЧАС, МПа/час» |
| press_ch_d | Средняя скорость изменения давления в коллекторе ИУ в СУТ, МПа/час» |
| press_ch_w | Средняя скорость изменения давления в коллекторе ИУ в НЕДЕЛЯ, МПа/час» |
| press_coll | Давление в коллекторе измерительной установки, МПа |
| workload_eng | Загрузка двигателя, % |
| curr | Ток фазы А двигателя, А |
| power_coef | Коэффициент мощности (cos fi) |
| liq | Объём жидкости в рабочих условиях за время наработки суточный, м3 |

Дополнительно, как признаки были выделены из индекса минуты, дни недели, и час, соответствующий временной метке для данного измерения. Это было сделано чтобы избежать возможную потерю информации, т.к. использованные библиотеки (Pycaret, PYOD)[1] не рассматривают индекс как параметр.

Таблица №2. Дополнительные параметры датасета

| **Название параметра** | **Описание** |
|---|---|
| hour | Час |
| day_w | День недели |
| minute | минута |

Рассмотрим признаки на примере отдельной насосной установки id = 226000188. Построим гистограмму распределения признаков (Рис. 1) и диаграмму корреляции (Рис. 2). Как мы видим, после того как технический

специалист провалидирует аномалии, для уменьшения количества обрабатываемых данных можно исключить признаки, которые сильно коррелируют с другими. В данной работе мы этого делать не будем, т.к. мы решаем задачу обучения без учителя, и не можем менять параметры датасета, т.к. нет однозначных метрик качества модели.


**Рис. 1.** Гистограмма распределения параметров для насоса id = 226000188

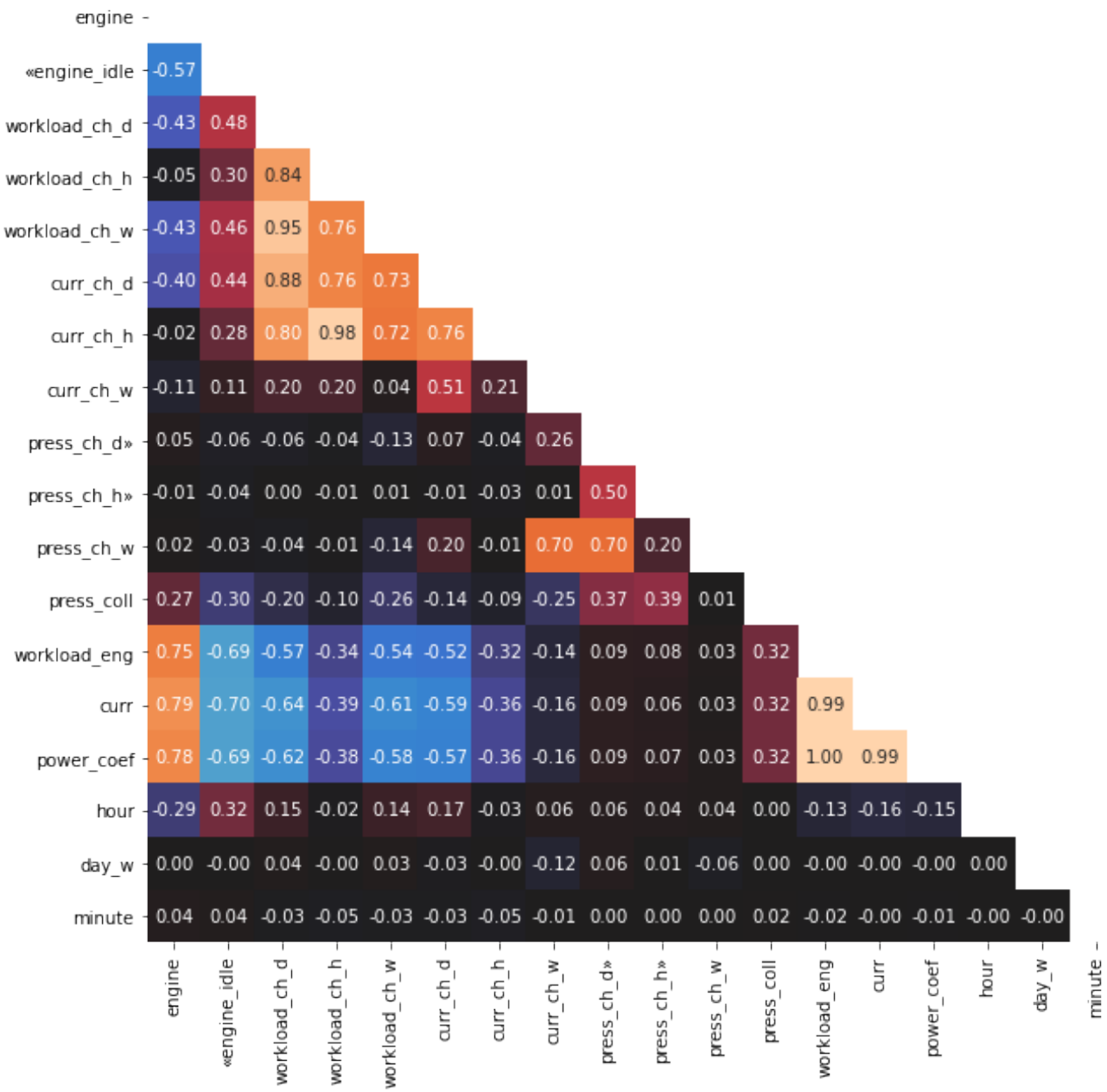



**Рис. 2.** Корреляционная матрица параметров для насоса id = 226000188, чем темнее оттенок, тем меньше коэффициент корреляции.

## Раздел III. Модели поиска аномалий

### Модель 1. k-Nearest Neighbors (k-NN)

Алгоритм, изначально использующийся для автоматической классификации объектов, также может применяться для поиска аномалий. В таком случае перед применением необходимо определить функцию расстояния (метрику). Классический вариантом такой функции является евклидово расстояние. Предполагается что у возможной аномалии расстояние до ближайших К-соседей будет выше[2].

### Модель 2. Isolation forest (i-forest)

Один из вариантов случайного леса (строится из решающих деревьев). Выбирается случайный признак и случайное разделение, по которым строится ветвление в дереве. Для каждого объекта выборки определяется мера его нормальности как среднее арифметическое глубин листьев, в которые он попал (под этим понимается ”изоляция”). При таком способе построения деревьев аномалии будут попадать в листья на ранних этапах (на небольшой глубине дерева), то есть выбросы проще ”изолировать”. Дерево строится до тех пор, пока каждый объект не окажется в отдельном листе). На рисунке 3 продемонстрирована основная идея метода[3].

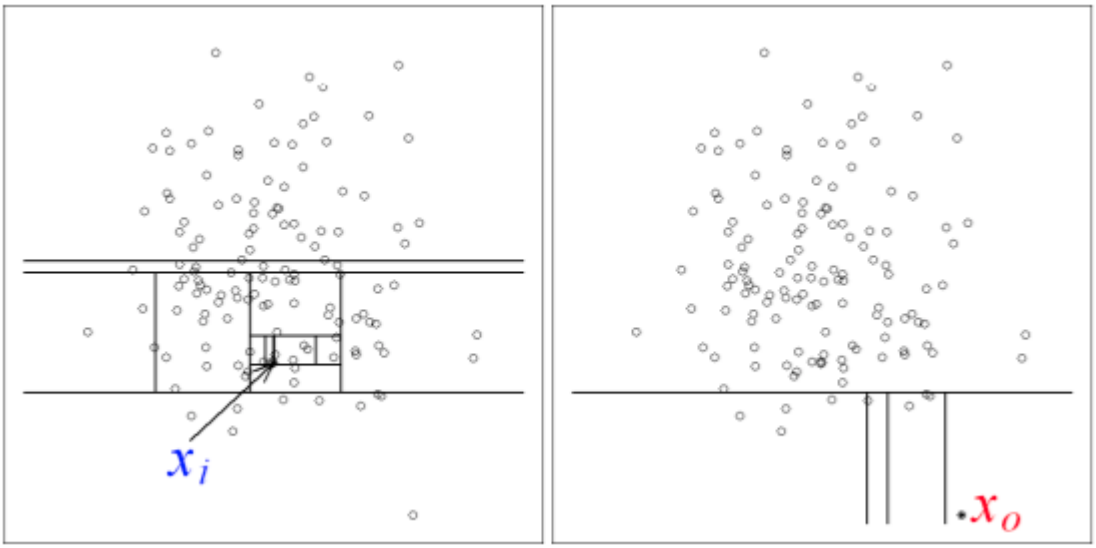



**Рис.3.** Для изолирования точки Xi требуется 12 случайных разбиений, а для аномальной точки Xo – только 4 разбиения.

## Модель 3. Principal Component Analysis (PCA)

Название алгоритма переводится как “метод главных компонент”. Он применяется во многих областях, в том числе, биоинформатике, обработке изображений, для сжатия данных, в общественных науках. Вычисление этих главных 16 компонент может быть сведено к вычислению сингулярного разложения матрицы данных или к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. На рисунке 4 представлены собственные векторы в случае двумерной гауссианы[4].

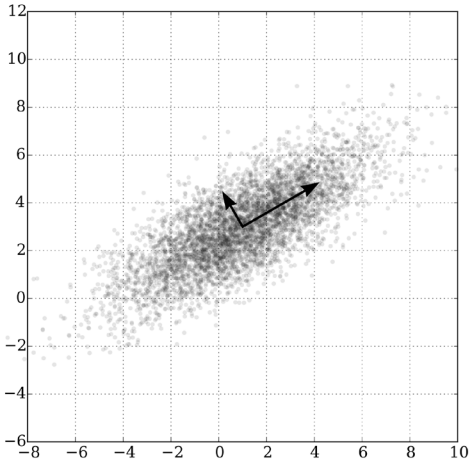



**Рис. 4.** PCA для многомерного гауссового распределения с центром в точке (1, 3) со стандартным отклонением 3. Векторы отражают собственные векторы ковариационной матрицы гауссианы.

### Модель 4. One-class SVM detector (SVM)

SVM - базовая линейная модель. Основная идея алгоритма (в случае с классификацией) - разделить классы гиперплоскостью так, чтобы максимизировать расстояние (зазор) между ними. Изначально алгоритм был способен работать только с линейно разделимыми классами, однако в 90-е годы прошлого века метод стал особенно популярен из-за внедрения "Kernel Trick" (1992), позволившего эффективно работать с линейно неразделимыми данными.

Ядро (kernel) - это функция, которая способна преобразовать признаковое пространство (в том числе нелинейно), без непосредственного преобразования признаков.Крайне эффективна в плане вычисления и потенциально позволяет получать бесконечноразмерные признаковые пространства[5].

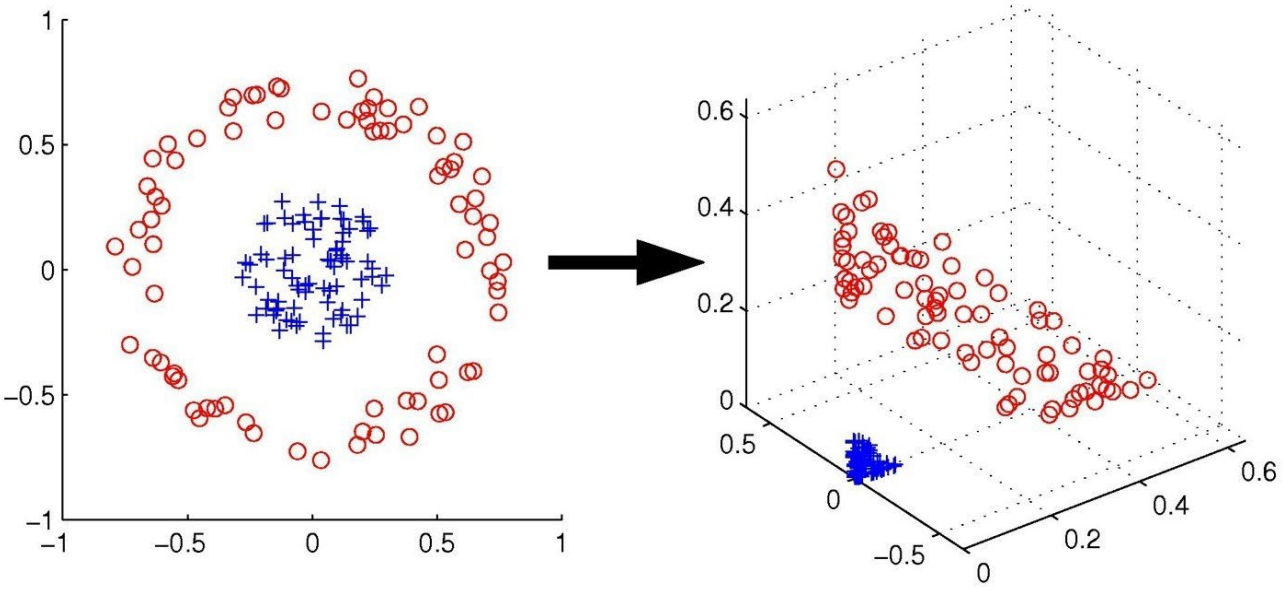



**Рис. 5.** Идея заключается в том, что классы, линейно неразделимые в текущем признаковом пространстве, могут стать разделимыми в пространствах более высокой размерности

One Class SVM – это одна из форм классического алгоритма, однако, как следует из названия, для его обучения нам достаточно иметь всего один класс, пусть даже

и немного "зашумленный", при этом мы хотим научиться для каждого наблюдения принимать решение, является ли оно аномальным или нет.

Общая идея: преобразовать признаковое пространство и провести разделяющую гиперплоскость так, чтобы наблюдения лежали как можно дальше от начала координат. В результате мы получаем границу, по одну сторону которой лежат максимально "плотные" и похожие друг на друга наблюдения из нашей тренировочной выборки, а по другую будут находится аномальные значения, не похожие на все остальные.

### Модель 5. Local Outlier Factor (LOF)

Метод Локального уровня выброса основывается на концепции локальной плотности, где локальность задаётся k-ближайшими соседями, расстояния до которых используются для оценки плотности. Путём сравнения локальной плотности объекта с локальной плотностью его соседей, можно выделить области с аналогичной плотностью и точки, которые имеют существенно меньшую плотность, чем её соседи. Эти точки считаются аномалиями. Локальная плотность оценивается расстоянием, с которым точка может быть “достигнута” от соседних точек. Определение “расстояния достижимости” используемого в алгоритме, является дополнительной мерой для получения более устойчивых результатов внутри кластеров[6].

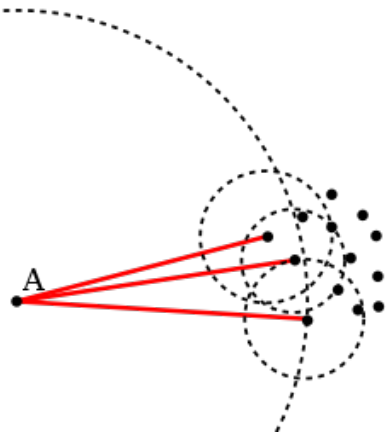



**Рис. 6.** На рисунке изображена точка с радиусом, в рамках которого для каждой точки измеряется плотность(количество) соседних точек

## Модель 6. Connectivity-Based Outlier Factor (COF)

COF является доработкой метода LOF путем изменения способа подсчета плотности окружающих точек. LOF плохо разделяет аномалии от паттернов с низкой плотностью (пример: точки распределенные по прямой и рядом лежащая аномалия).

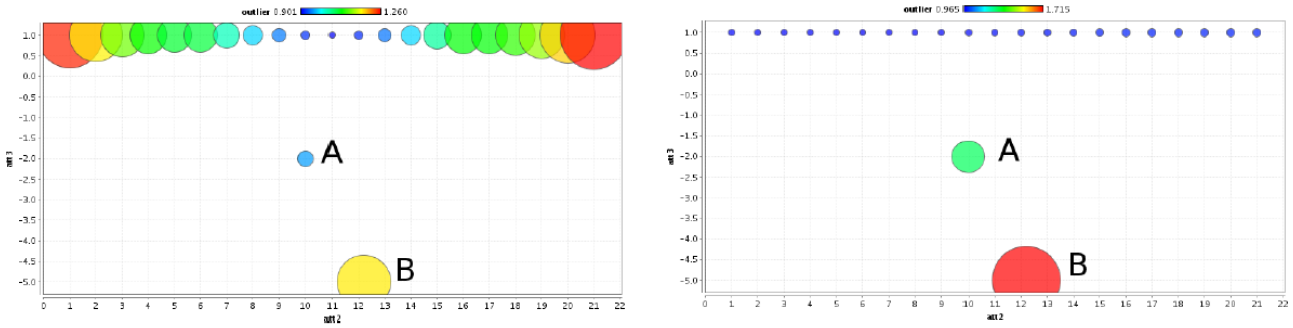



**Рис. 7.** На рисунке слева изображены результаты классификации LOF, справа COF (красная точка – более высокая степень аномалии, синяя – низкая)

Метод COF считает расстояние не по прямой в рамках радиуса, а вводит понятие цепочного расстояния, которое представляет подсчет количества граней графа, который последовательно соединяет K-ближайших точек[6]. Общий принцип измененного подхода модели к учету расстояния проиллюстрирован на рисунке 8.

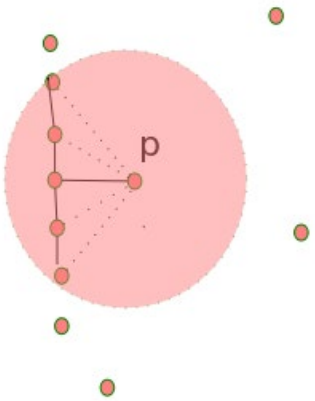



**Рис. 8.** На рисунке изображены различные подходы к подсчету расстояния: закрашенный круг – радиус доступности, в рамках которого мы считаем точки (плотность), пунктиром – подход модели LOF, сплошные линии – COF.

## Модель 7. Clustering-Based Local Outlier (CBLOF)

Метод уровня выброса, основанного на кластеризации (Clustering-Based Local Outlier - CBLOF) - был разработан в первой половине 2000х с целью получить вычислительно менее трудоемкий и более чувствительный, по сравнению с уже имеющимися на тот момент (ROCK, $C^2P$, DBSCAN) алгоритм. В данном методе мы полученные данные разбиваем на кластеры (можно разбивать различными методами, в используемой в данной работе алгоритме библиотеки PYOD – на основании плотности К-средних), далее на основании заданных как гиперпараметры α- и β- коэффициентов, кластеры ранжируются от больших к маленьким, и затем для точек считаем фактор аномальности на основании размера кластера, к которому принадлежит точка, и расстояния от данной точки до центра ближайшего большого кластера [7].

Интуитивно алгоритм проиллюстрирован ниже на рисунке 9. Общая трудоемкость алгоритма O(N).

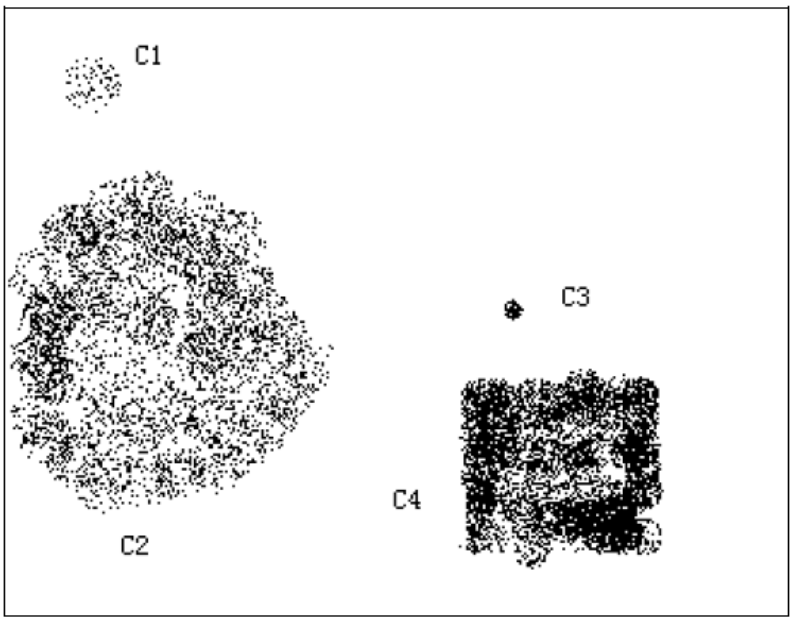



**Рис. 9.** Интуитивная иллюстрация к методу CBLOF. C2 и C4 относятся к кластеру большого размера, а C1 и C3 относятся к кластерам малого размера и их “счет выброса(аномальности)” будет выше.

### Модель 8. Histogram-based Outlier Detection (HBOS)

Алгоритм HBOS в несколько раз быстрее работает, чем алгоритмы, основанные на кластеризации и методе ближайших соседей. Для каждого измерения d строится одномерная гистограмма, где высота каждой ячейки отражает оценку плотности. Затем проводится нормировка таким образом, что максимальная высота ячеек каждой гистограммы составляет 1. Это обеспечивает равный вклад каждого измерения в оценку аномальности. Наконец, для каждого объекта $p$ выборки рассчитывается HBOS, используя высоту соответствующей ячейки, в которой объект расположен:

$$HBOS\ (p)\ = \sum_{i=0}^{d} log\left(\frac{1}{hist_i(p)}\right)$$

Вместо произведения используется сумма логарифмов – это то же самое, что и применение логарифма к произведению (log (a · b) = log (a) + log (b)). Такой подход менее чувствительный к ошибкам, которые связаны с точностью плавающей точки в экстремально несбалансированных распределениях, что в свою очередь может приводить к очень высоким значениям оценки аномальности [8].

### Модель 9. Subspace Outlier Detection (SOD)

Общая идея метода SOD - найти множество референсных точек, например, ближайших соседей. Затем используя подмножество признаков создать гиперплоскость, вдоль которой лежит максимальное количество признаков. И затем посмотреть, как далеко лежит от этой плоскости лежит наша точка [9].

Интуитивно алгоритм проиллюстрирован ниже на рисунке 10**.** Общая трудоемкость алгоритма O($d \cdot n^2$), где d – размерность датасета.

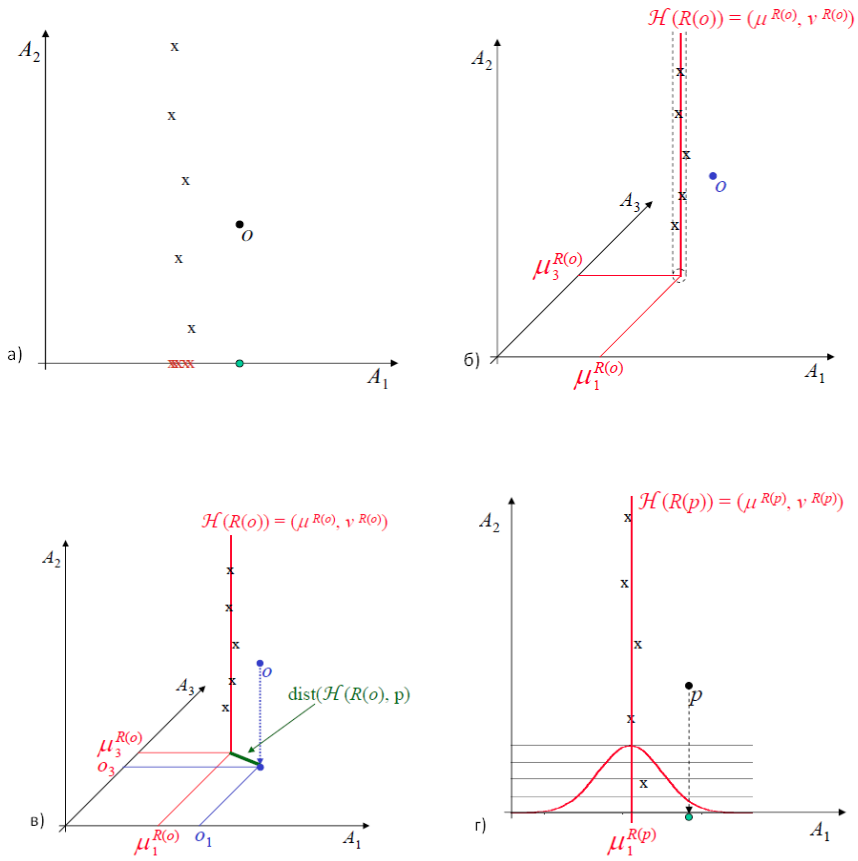



**Рис. 10.** Интуитивная иллюстрация к методу SOD: а) т. O является аномалией, но при этом локальная плотность вокруг данной точки не отличается от соседних точек; б) находим линию вдоль которой расположено большинство точек подмножества наблюдений; в) проводим через данную прямую и т. O гиперплоскость; в) строим график распределения расстояний в гиперплоскости от прямой до точки, в примере на рисунке т. O находится на конце нормального распределения случайной величины, что свидетельствует о высокой вероятности того, что т. O – аномалия.

## Модель 10. Angle-base Outlier Detection (ABOD)

Метод определения аномалий на основании углов – метод, в котором оценка аномальности точки считается не на основании расстояния до ближайших K-соседей, а на основании спектра углов между векторами, соединяющими точку и две точки из K-соседей. Общая идея заключается в том, что углы между векторами точек кластера будут сильно варьироваться, в тоже время как для аномалии угол будет изменяться незначительно.

Интуитивно алгоритм проиллюстрирован ниже на рисунке 11. Общая трудоемкость алгоритма O($n^3$) [10].

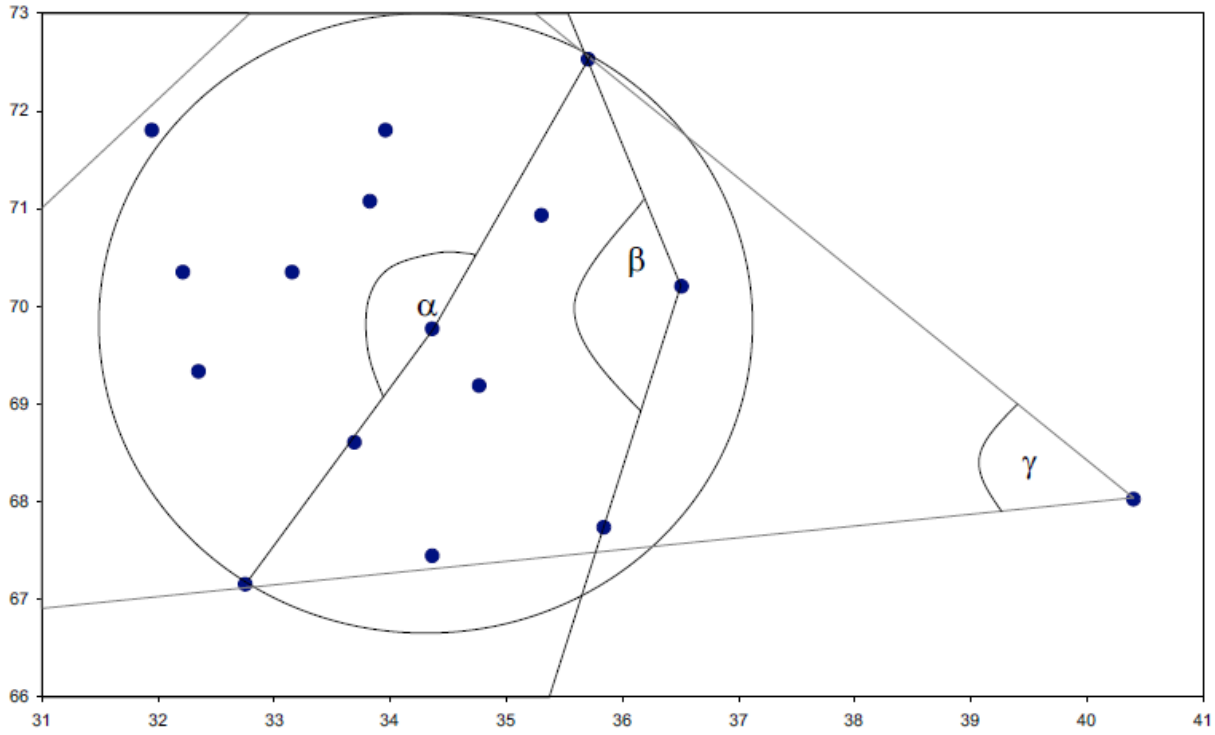



**Рис. 11.** Интуитивная иллюстрация к методу ABOD: углы (α, β) между векторами соединяющие точку наблюдения и соседние точки, находящиеся в кластере будут сильно отличаться между собой, в тоже время для точки лежащей вне кластера, какие бы две другие точки мы не взяли, угол (γ) не будет сильно разниться.

## Модель 11. Minimum Covariance Determinant (MCD)

Метод минимизации ковариации детерминанта – устойчивый(робастный) метод определения аномалий. Его основная идея заключается в том, что мы предполагаем, что совокупность признаков подчинена многомерному распределению (нормальное), определяется центр распределения и разброс, при помощи вектора средних и ковариционной матрицы. И точки, наиболее далеко расположенные от центра этого распределения с большей степенью вероятности являются аномалиями. Расстояние, по которому считается показатель аномалии – расстояние Махаланобиса [11].

Расстояние Махаланобиса между двумя точками — мера расстояния между двумя случайными точками U и V, одна из которых может (или обе могут) принадлежать некоторому классу C с матрицей ковариаций COV:

$$d_M(U,V,COV^{-1}) = \sqrt{(U-V)\cdot COV^{-1}\cdot(U-V)^T}$$

Символ T означает операцию транспонирования, а под $COV^{-1}$ подразумевается матрица, обратная ковариационной.

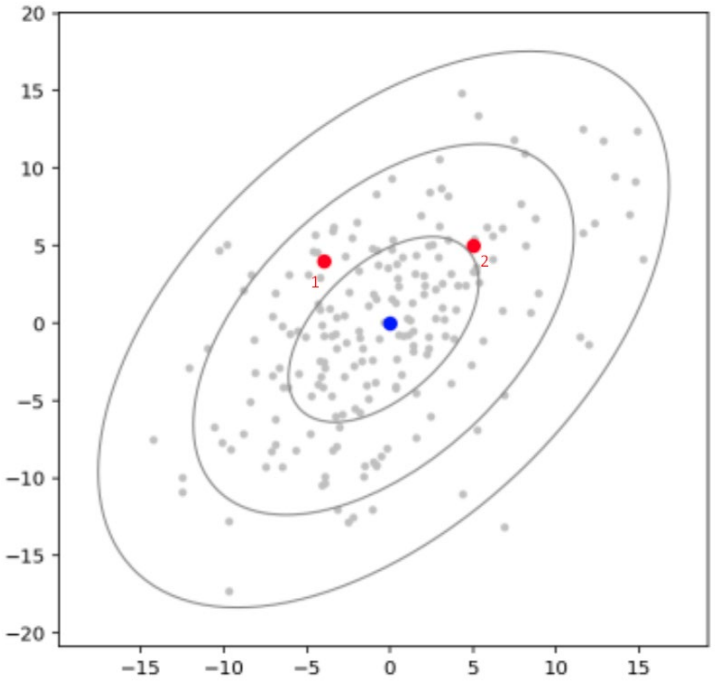



**Рис. 12.** Интуитивная иллюстрация к методу учета расстояния в методе MCD: точка 1, если считать Евклидово расстояние, находится ближе к центру распределения(синяя точка) чем точка 2. В то же время, расстояние Махаланобиса, учитывающее дисперсию и ковариацию, меньше у точки 2.

## Раздел IV. Обработка данных

### Результат применения различных моделей на примере одно насоса

Для иллюстрирования итогов применения моделей поиска аномалий спроецируем данные для одного насоса (id = 226000188) при помощи метода t-SNE на двумерную плоскость.

Стохастическое вложение соседей с t-распределением (англ. t-Distributed

Stochastic Neighbor Embedding, t-SNE) — метод визуализации данных высокой размерности с помощью представления каждой точки данных в двух или трехмерном пространстве. Ниже представлен ряд диаграмм с визуализацией параметров для одного насоса. Каждая группа точек представляет собой отдельный режим работы. Желтым цветом обозначены точки, являющиеся аномалиями по итогам применения соответствующего алгоритма.

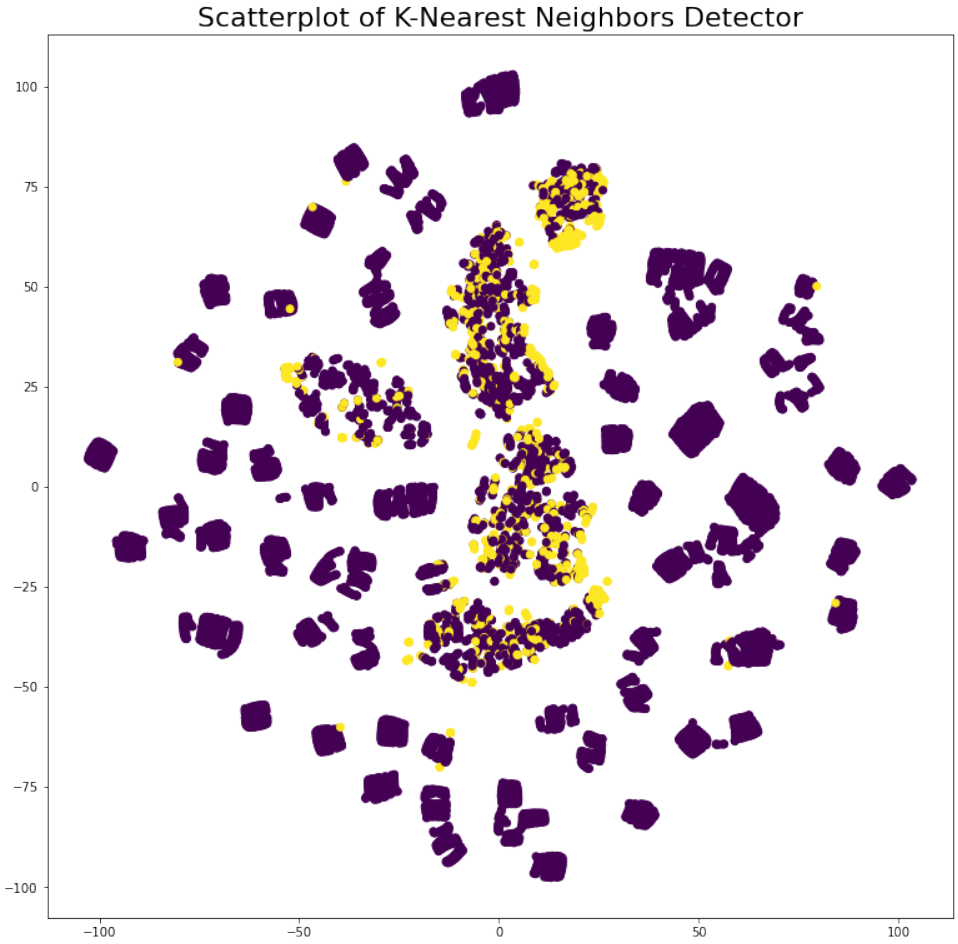



**Рис. 13.** Результат применения модели **k-NN** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

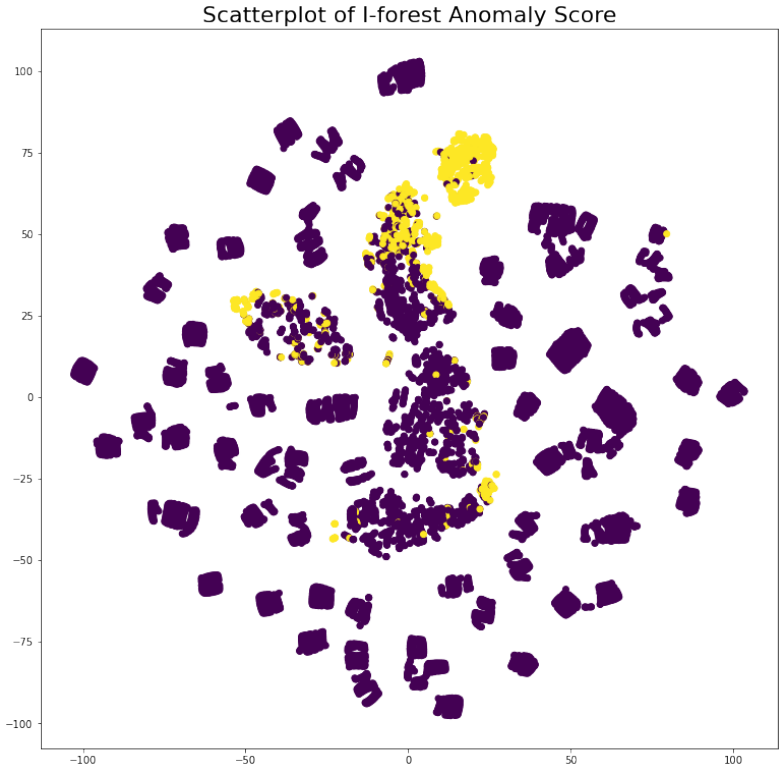



**Рис. 14.** Результат применения модели **i-forest** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

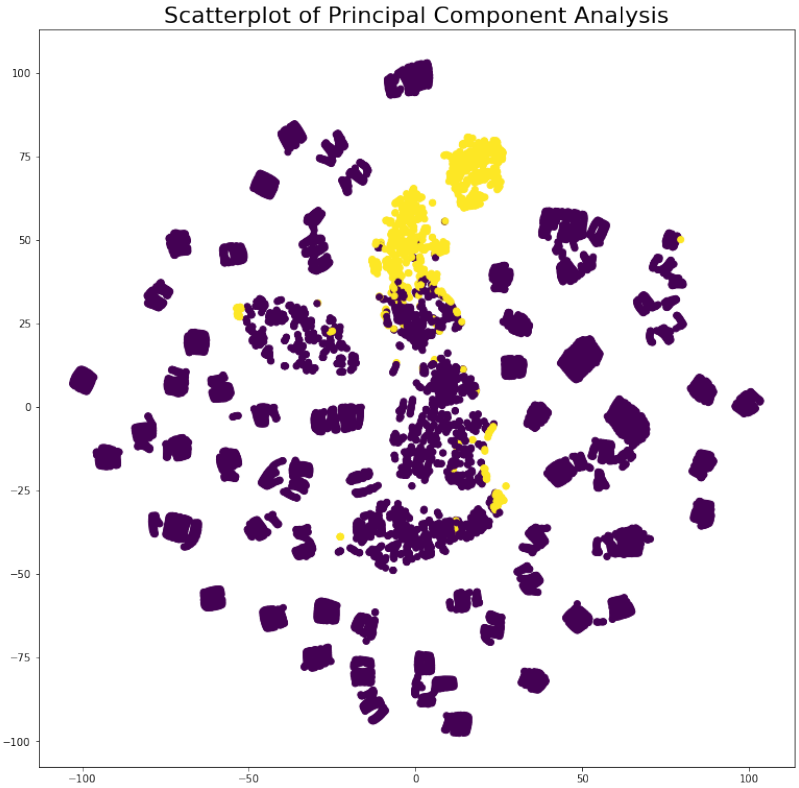



**Рис. 15.** Результат применения модели **PCA** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

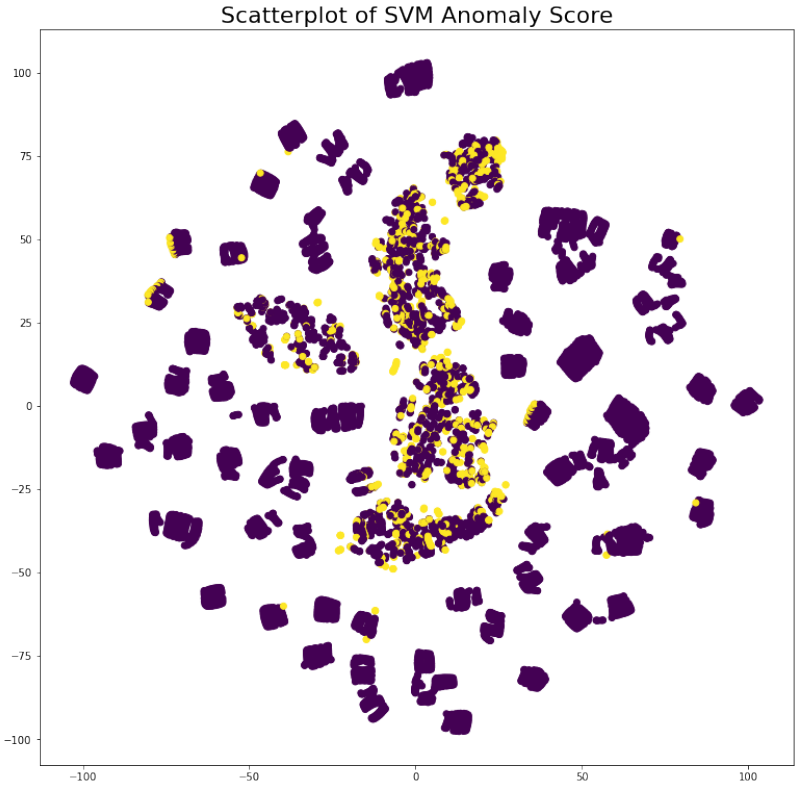



**Рис. 16.** Результат применения модели **SVM** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

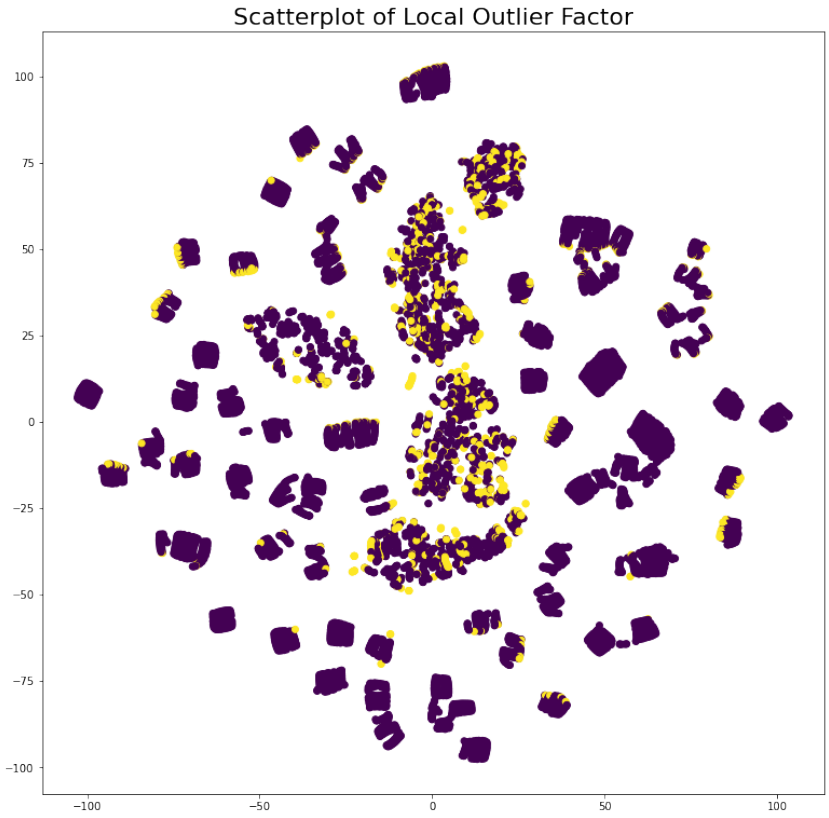



**Рис. 17.** Результат применения модели **LOF** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

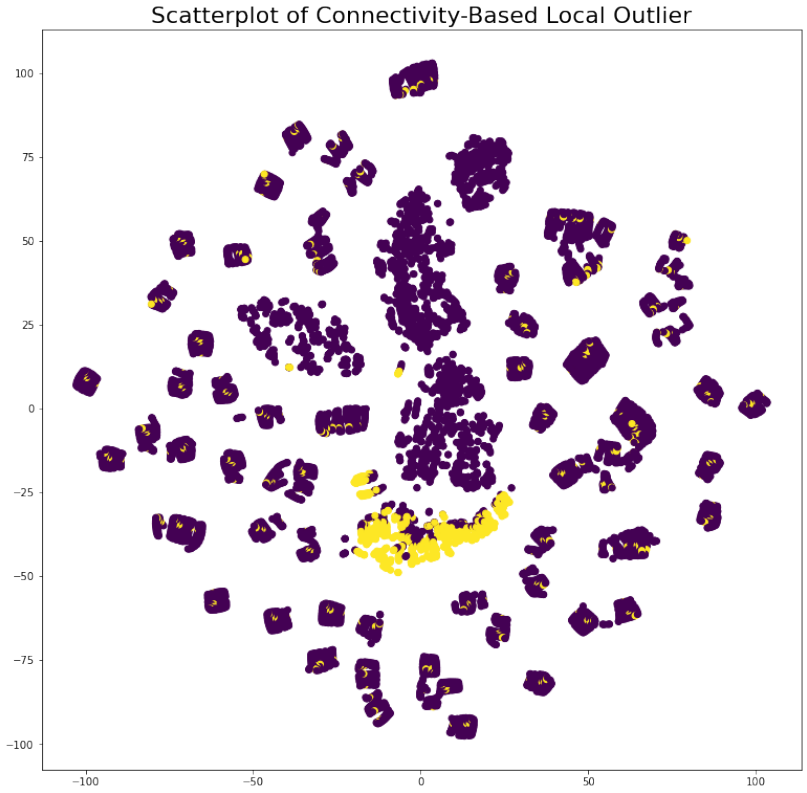



**Рис. 18.** Результат применения модели **COF** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

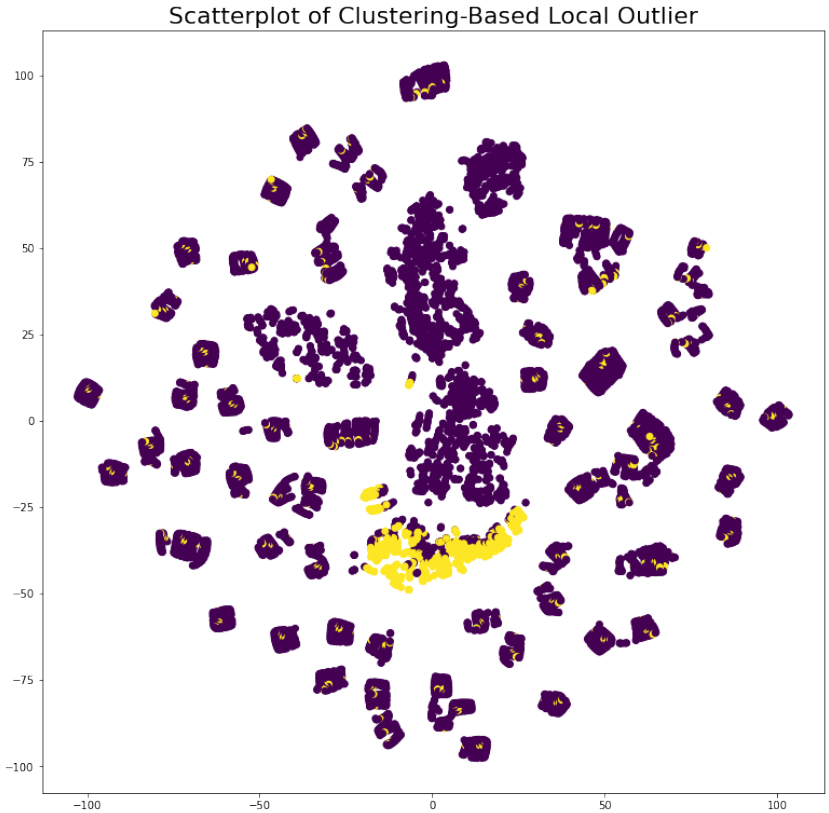



**Рис. 19.** Результат применения модели **CBLOF** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

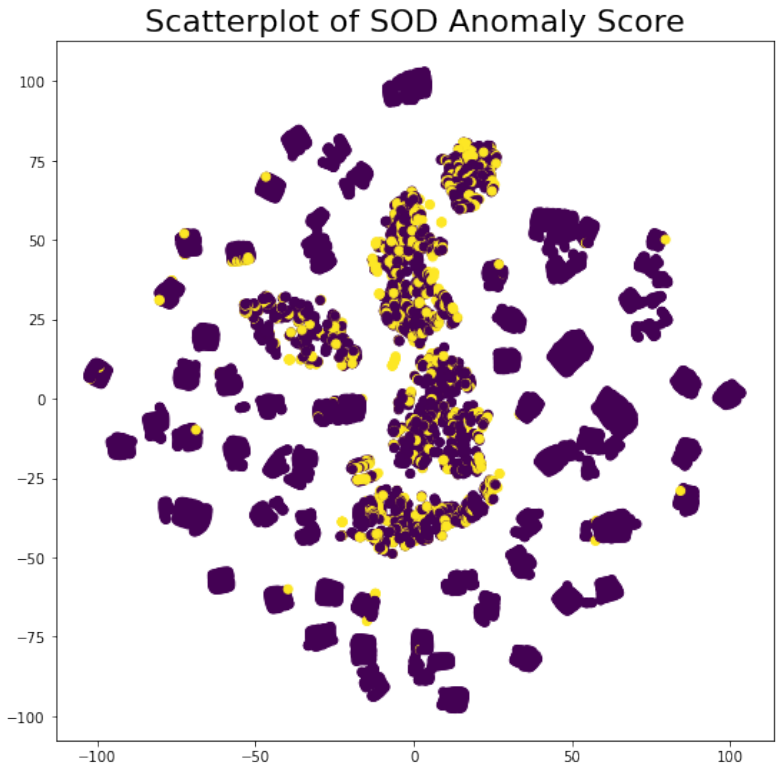



**Рис. 20.** Результат применения модели **SOD** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

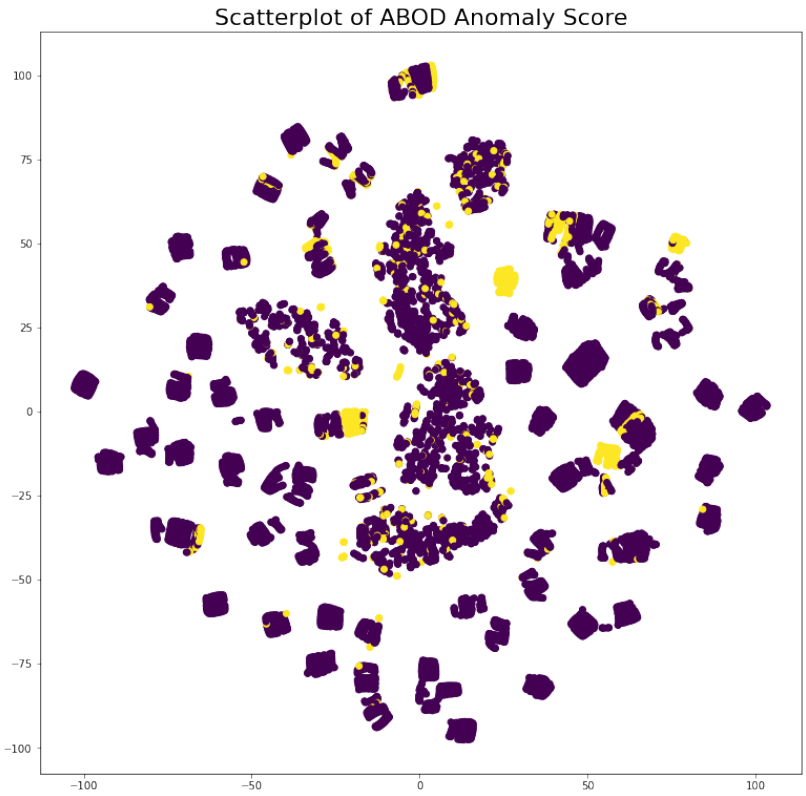



**Рис. 21.** Результат применения модели **ABOD** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

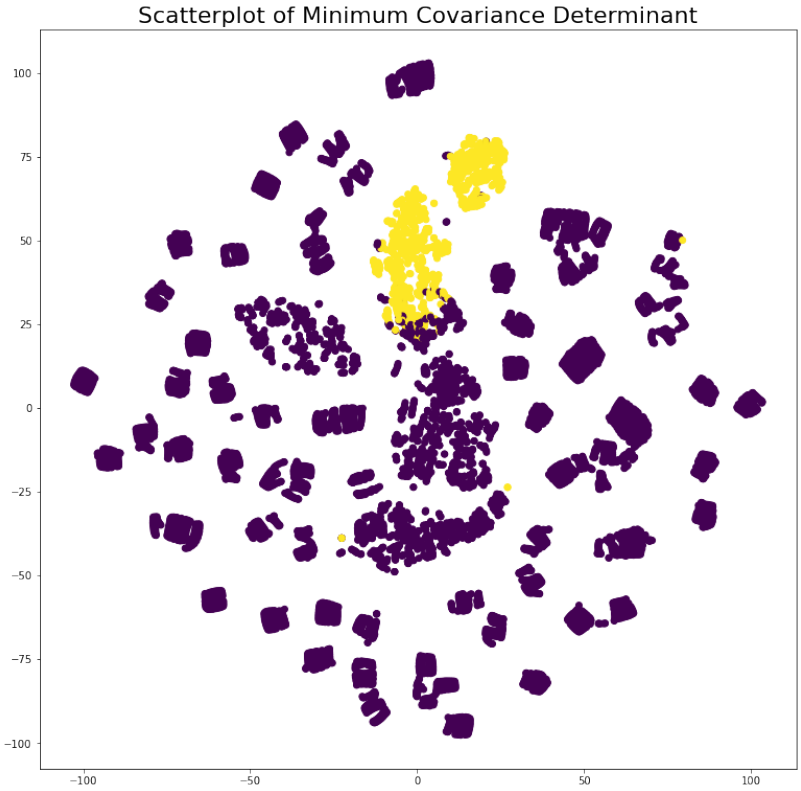



**Рис. 22.** Результат применения модели **MCD** (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

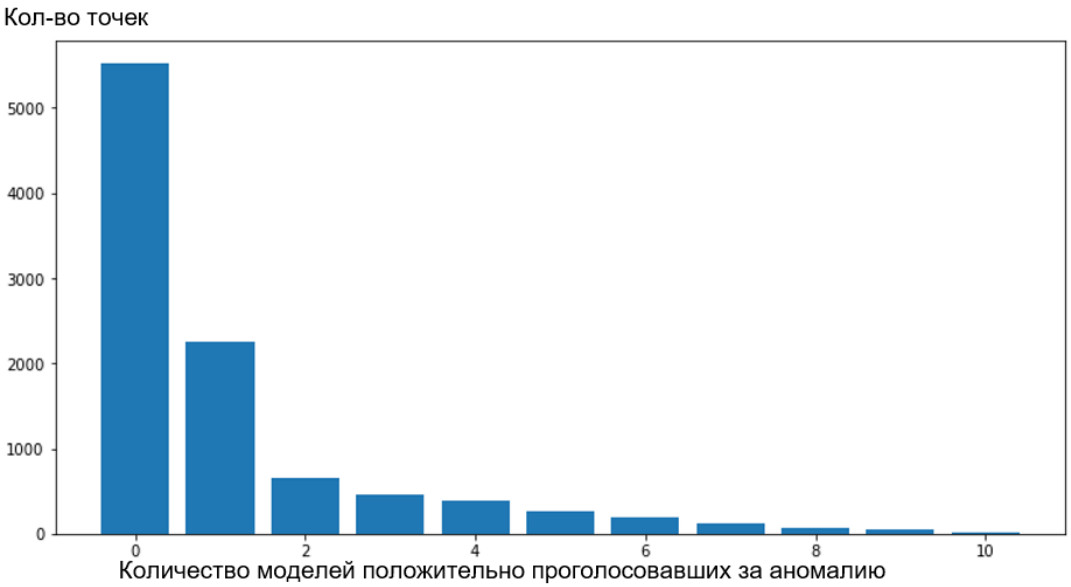
## Ансамблирование различных моделей на примере одно насоса

По итогам применения различных методов поиска аномалий мы получаем итоговую сводную таблицу размером 9975x11 (9975 – количество наблюдений для насоса с id = 226000188), где для каждого наблюдения в соответствующем столбце для каждого метода будет указано значение 0 или 1 (аномалия данная точка или нет).

Просуммировав количество голосов для каждой точки, мы получим гистограмму ансамблирования методов, где в качестве горизонтальной оси будет выступать количество положительно проголосовавших моделей, а по вертикали – количество точек, за которое одновременно, проголосовало указанное количество моделей.

Таблица №3. Фрагмент сводной таблицы голосования моделей по данным одного насоса

| | IFOREST | ABOD | CBLOF | COF | HBOS | KNN | LOF | SVM | PCA | MCD | SOD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2019-06-27 06:40:00** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2019-06-27 06:45:00** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2019-06-27 06:50:00** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2019-06-27 06:55:00** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |




| модели | точки |
|---|---|
| 0 | 5521 |
| 1 | 2245 |
| 2 | 653 |
| 3 | 453 |
| 4 | 392 |
| 5 | 269 |
| 6 | 185 |
| 7 | 120 |
| 8 | 75 |
| 9 | 43 |
| 10 | 19 |
| 11 | 0 |

**Рис. 23.** Гистограмма ансамблирования методов поиска аномалий для данных по одному насосу (id = 226000188).

Общий смысл ансамблирования заключается в следующем – чем больше моделей положительно проголосовало за точку как за аномалию, тем с большей уверенностью будем можно считать данную точку аномальным значением. Проиллюстрируем аномальные точки на t-SNE проекции параметров исходя из

того, что аномалией мы будем считать точку, за которую проголосовало N= 5, N= 8, N= 10 моделей соответственно.

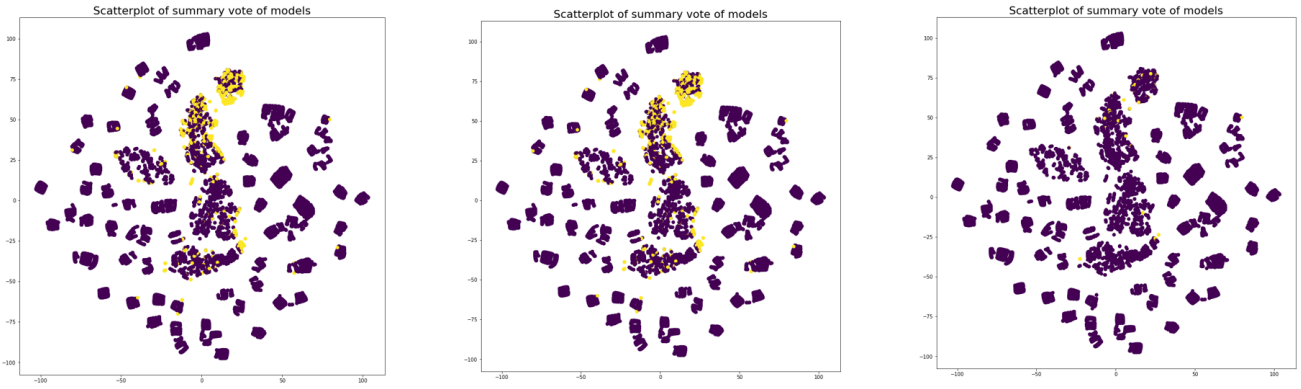



**Рис. 24.** Результат применения ансамблирования методов (точки выделенные желтым – аномалия) на t-SNE проекции данных по одному насосу (id = 226000188).

Принимая в расчет тот факт, что все охарактеризованные как аномалии точки до их валидации техническим специалистом являются только предварительными результатами, мы можем дать только предварительную и относительную оценку моделей, сравнив их между собой.

Для этого предположим, что те точки, за которые проголосовало N=5 или больше моделей, являются аномалиями, а те, за которые меньше – не являются аномалиями. Посчитаем метрики для каждой модели по по одному насосу (id = 226000188).

Таблица №4. Метрики качества моделей, при ансамблировании N=5 моделей(модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **K-NN** | **0.942356** | **0.797468** | **0.663546** | **0.568136** |
| **PCA** | **0.939950** | **0.780591** | **0.649503** | **0.556112** |

| | | | | |
|---|---|---|---|---|
| **HBOS** | **0.933935** | **0.738397** | **0.614394** | **0.526052** |
| I-forest | 0.932932 | 0.731364 | 0.608543 | 0.521042 |
| SVM | 0.927519 | 0.693390 | 0.576946 | 0.493988 |
| MCD | 0.925915 | 0.682138 | 0.567583 | 0.485972 |
| LOF | 0.915890 | 0.611814 | 0.509070 | 0.435872 |
| SOD | 0.909273 | 0.565401 | 0.470451 | 0.402806 |
| ABOD | 0.876792 | 0.337553 | 0.280866 | 0.240481 |
| COF | 0.865564 | 0.258790 | 0.215331 | 0.184369 |
| CBLOF | 0.844712 | 0.092827 | 0.078525 | 0.068041 |

Т.к. классы в нашей задаче сильно не сбалансированы, то метрика accuracy - доля правильных ответов алгоритма не является информативной. Будем ориентироваться на метрики precision и recall, и косвенно на f1 меру, как на сочетание этих двух метрик. Precision можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а recall показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Для насоса (id = 226000188), если считать аномалиями за которые проголосовало N=5 или больше моделей лучшими относительно других примененных моделями являются: k-NN (recall_score = 0.797468), PCA (recall_score = 0.780591), HBOS(recall_score = 0.738397).

Составим аналогичные таблицы метрик, если считать аномалиями точки, за которые проголосовало N= 8, N=10 моделей соответственно.

Таблица №5. Метрики качества моделей, при ансамблировании N=8 моделей(модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **K-NN** | **0.924311** | **0.972763** | **0.398406** | **0.250501** |
| **SVM** | **0.923910** | **0.964981** | **0.395219** | **0.248497** |
| **LOF** | **0.920702** | **0.902724** | **0.369721** | **0.232465** |
| PCA | 0.919699 | 0.883268 | 0.361753 | 0.227455 |
| I-forest | 0.917694 | 0.844358 | 0.345817 | 0.217435 |
| HBOS | 0.914085 | 0.774319 | 0.317131 | 0.199399 |
| MCD | 0.910276 | 0.700389 | 0.286853 | 0.180361 |
| SOD | 0.908471 | 0.665370 | 0.272510 | 0.171343 |
| ABOD | 0.903058 | 0.560311 | 0.229482 | 0.144289 |
| COF | 0.897644 | 0.455253 | 0.186454 | 0.117234 |
| CBLOF | 0.883409 | 0.124514 | 0.052160 | 0.032990 |

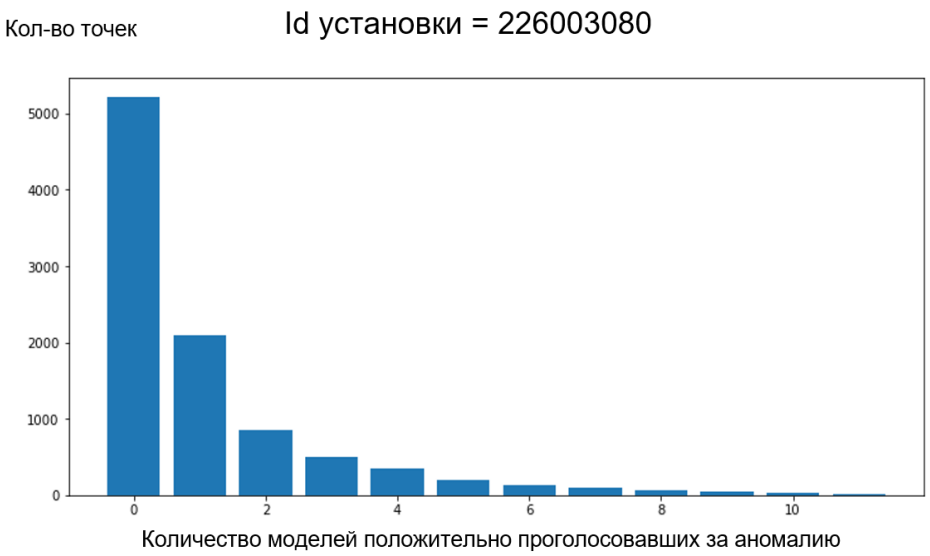
Таблица №6. Метрики качества моделей, при ансамблировании N=10 моделей (модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **I-forest** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **ABOD** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |

| | | | | |
|---|---|---|---|---|
| **COF** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **k-NN** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **LOF** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **SVM** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **PCA** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| **SOD** | **0.901855** | **1.000000** | **0.037365** | **0.019038** |
| HBOS | 0.901654 | 0.947368 | 0.035398 | 0.018036 |
| MCD | 0.901454 | 0.894737 | 0.033432 | 0.017034 |
| CBLOF | 0.901454 | 0.157895 | 0.006067 | 0.003093 |

Как видно по таблице 5 – лучший результат все так же показывает модель k-NN (recall_score = 0.972763), затем PCA (recall_score = 0.964981), и LOF(recall_score = 0.902724). По таблице 6 видно, что в случае голосования 10 моделями метрики теряют смысл и информативность. Поэтому для относительно сравнения по остальным насосным установкам будем использовать подход, в котором аномалией считается модель, за которую проголосовало N= 5 или больше моделей.

На рисунке N изображены гистограммы ансамблирования по трем другим насосам из датасета как иллюстрация распределения точек по количеству проголосовавших моделей.




| модели | точки |
|---|---|
| 0 | 5214 |
| 1 | 2102 |
| 2 | 853 |
| 3 | 500 |
| 4 | 355 |
| 5 | 196 |
| 6 | 127 |
| 8 | 89 |
| 7 | 71 |
| 9 | 55 |
| 10 | 32 |
| 11 | 6 |


| модели | точки |
|---|---|
| 0 | 4958 |
| 1 | 1598 |
| 3 | 719 |
| 5 | 472 |
| 2 | 446 |
| 4 | 394 |
| 6 | 71 |
| 9 | 37 |
| 8 | 13 |
| 10 | 11 |
| 7 | 5 |
| 11 | 1 |


| модели | точки |
|---|---|
| 0 | 5537 |
| 1 | 2299 |
| 2 | 609 |
| 3 | 423 |
| 4 | 358 |
| 5 | 288 |
| 6 | 184 |
| 7 | 114 |
| 8 | 84 |
| 9 | 49 |
| 10 | 27 |
| 11 | 3 |

**Рис. 25.** Гистограмма ансамблирования методов поиска для данных по одному насосу (взято 3 произвольных насоса из 17).

## Ансамблирование различных моделей для всех установок датасета

Для того, чтобы сделать выводы о сравнительной эффективности моделей посчитаем метрики качества моделей по всем насосам, а затем усредним метрики.

Таблица №7. Метрики качества моделей по всем насосным установкам, при ансамблировании N=5 моделей (модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **k-NN** | **0.940024** | **0.798447** | **0.627838** | **0.529481** |
| **SVM** | **0.928465** | **0.706016** | **0.557811** | **0.471711** |
| **PCA** | **0.926297** | **0.700766** | **0.548282** | **0.460870** |
| I-forest | 0.921623 | 0.662130 | 0.519861 | 0.437501 |
| HBOS | 0.921262 | 0.664575 | 0.519285 | 0.435693 |
| MCD | 0.913973 | 0.602470 | 0.473293 | 0.399268 |
| SOD | 0.907905 | 0.562947 | 0.439294 | 0.368943 |
| LOF | 0.904274 | 0.525368 | 0.414058 | 0.350787 |
| COF | 0.892047 | 0.443305 | 0.343991 | 0.289667 |
| ABOD | 0.876727 | 0.295276 | 0.245014 | 0.213081 |
| CBLOF | 0.874128 | 0.288201 | 0.231884 | 0.197118 |

Таблица №8. Метрики качества моделей по всем насосным установкам, при ансамблировании N=8 моделей (модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **k-NN** | **0.912692** | **0.980204** | **0.218371** | **0.129066** |
| **SVM** | **0.912690** | **0.974932** | **0.218673** | **0.129295** |
| **PCA** | **0.911772** | **0.963185** | **0.211590** | **0.124702** |
| I-forest | 0.911136 | 0.930469 | 0.205980 | 0.121523 |
| HBOS | 0.910416 | 0.929303 | 0.200275 | 0.117835 |
| LOF | 0.908975 | 0.850900 | 0.187241 | 0.110723 |
| MCD | 0.908943 | 0.797534 | 0.187258 | 0.110562 |
| SOD | 0.906140 | 0.758163 | 0.163953 | 0.096557 |
| COF | 0.905613 | 0.710605 | 0.158087 | 0.093915 |
| ABOD | 0.901068 | 0.447457 | 0.116536 | 0.071184 |
| CBLOF | 0.896058 | 0.260001 | 0.070137 | 0.041713 |

Таблица №9. Метрики качества моделей по всем насосным установкам, при ансамблировании N=5 моделей (модели расположены по убыванию метрик, выше - лучше).

| Имя модели/метрики | accuracy_score | recall_score | f1_score | precision_score |
|---|---|---|---|---|
| **k-NN** | **0.940024** | **0.798447** | **0.627838** | **0.529481** |
| **SVM** | **0.928465** | **0.706016** | **0.557811** | **0.471711** |
| **PCA** | **0.926297** | **0.700766** | **0.548282** | **0.460870** |
| I-forest | 0.921623 | 0.662130 | 0.519861 | 0.437501 |
| HBOS | 0.921262 | 0.664575 | 0.519285 | 0.435693 |
| MCD | 0.913973 | 0.602470 | 0.473293 | 0.399268 |
| SOD | 0.907905 | 0.562947 | 0.439294 | 0.368943 |
| LOF | 0.904274 | 0.525368 | 0.414058 | 0.350787 |
| COF | 0.892047 | 0.443305 | 0.343991 | 0.289667 |
| ABOD | 0.876727 | 0.295276 | 0.245014 | 0.213081 |
| CBLOF | 0.874128 | 0.288201 | 0.231884 | 0.197118 |

По таблицам 7-9 можем отметить, что, вне зависимости от количества моделей ансамблирования (N= 5, N= 8, N= 10), **лучшие результаты показывают (по убыванию) модели: k-NN, SVM, PCA.**

## Раздел V. Выводы

1. На данных с насосного оборудования, используемого в нефтяной промышленности, были рассмотрены модели поиска аномалий, а также проведен их предварительный сравнительный анализ

2. Финальная оценка будет возможна по итогам подтверждения техническим специалистом предполагаемых аномалий

3. Лучшими моделями на представленном датасете предварительно показали себя модели:

    a. Метод KNN (K-Nearest Neighbors Detector)

    b. SVM (One-class SVM detector)

    c. PCA (Principal Component Analysis)

4. Данные модели, или их ансамбль может заменить стандартный подход поиска аномалии по каждому признаку отдельно по выходящим за интервал 3-сигма значениям

## Список литературы:


1. Документация библиотеки PYOD
   URL: https://pyod.readthedocs.io/en/latest/_modules/pyod/models/

2. Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner/ Mennatallah Amer and Markus Goldstein
   URL: https://www.goldiges.de/publications/Anomaly_Detection_Algorithms_for_RapidMiner.pdf

3. Isolation-based Anomaly Detection/ Fei Tony Liu and Kai Ming Ting, Zhi-Hua Zhou
   URL: https://www.researchgate.net/publication/239761771_Isolation-Based_Anomaly_Detection

4. A Tutorial on Principal Component Analysis/ Jonathon Shlens
   URL: https://arxiv.org/pdf/1404.1100.pdf

5. Anomaly detection using one-class SVM with wavelet packet decomposition/ Hannu Hautakangas, Jukka Nieminen
   URL: https://jyx.jyu.fi/bitstream/handle/123456789/37465/URN:NBN:fi:jyu-201202291321.pdf?sequence=1

6. Comparison of Unsupervised Anomaly Detection Techniques/ Mennatallah Amer
   URL: https://madm.dfki.de/_media/theses/thesis-amer.pdf

7. Discovering Claster-based Local Outliers/ Zengyou HE, Xiaofei Xu, Shengchun Deng
   URL: https://www.sciencedirect.com/science/article/abs/pii/S0167865503000035?via%3Dihub

8. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm/ Markus Goldstein, Andreas Dengel
   URL: https://www.researchgate.net/publication/231614824_Histogram-based_Outlier_Score_HBOS_A_fast_Unsupervised_Anomaly_Detection_Algorithm

9. Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data/ Hans-Peter Kriegel, Peer Kroger, Erich Schubert, Arthur Zimek
   URL: https://www.dbs.ifi.lmu.de/Publikationen/Papers/pakdd09_SOD.pdf

10. Angle-Based Outlier Detection in High-dimensional Data/ Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek
    URL: https://www.dbs.ifi.lmu.de/~zimek/publications/KDD2008/KDD08-ABOD.pdf

11. Minimum Covariance Determinant and Extensions/ Mia Hubert, Michiel Debruyne, Peter J. Rousseeuw
    URL: https://wis.kuleuven.be/stat/robust/papers/publications-2018/hubertdebruynerousseeuw-mcdext-wires-2018.pdf