Yandex

Обнаружение аномалий в данных телеметрии УЭЦН

Выполнил
Острик Петр

Руководитель
Абдуракипов Сергей

Специалист по Data Science
программа профессиональной переподготовки

Содержание

Постановка задачи

- **Оборудование** – электрические погружные насосы
- **Сфера применения** – нефтедобывающий сектор
- **Бизнес задача** – снизить количество выходящего из строя оборудования. Для этого необходимо определять аномальные показатели работы
- **Задача** – обнаружение аномалий в данных в многокомпонентных временных рядах (по множеству параметров одновременно)
- Не все насосы оборудованы телеметрическими датчиками на погруженной части
- Необходимо определять аномальные режимы на основании известных показателей установки, расположенной на поверхности

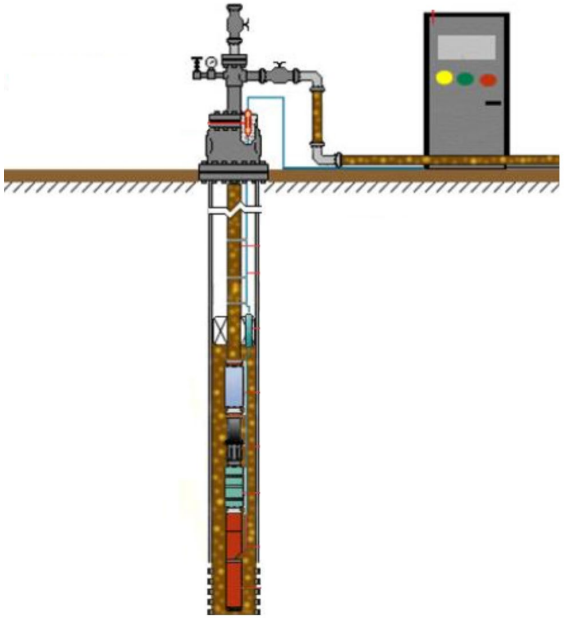


Рис.1 Общая схема погружного насоса

Обзор данных

Данные по 17 насосам за месяц (34 дня) с 2019-06 по 2019-07 с шагом в 5 мин

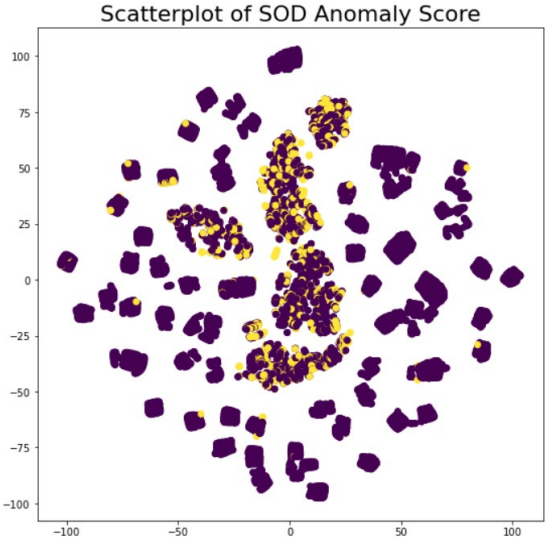
Параметры:

- Наработка двигателя с момента последнего включения, сек
- Время простоя двигателя с момента последнего выключения, сек
- Средняя скорость изменения загрузки двигателя, %/час
- Средняя скорость изменения тока фазы А двигателя в сутки, А/час
- Средняя скорость изменения давления в коллекторе в СУТ, МПа/час
- Давление в коллекторе измерительной установки, Мпа
- Загрузка двигателя, %
- Ток фазы А двигателя, А
- Коэффициент мощности
- Объём жидкости в рабочих условиях за время наработки суточный, м3

Модели поиска аномалий

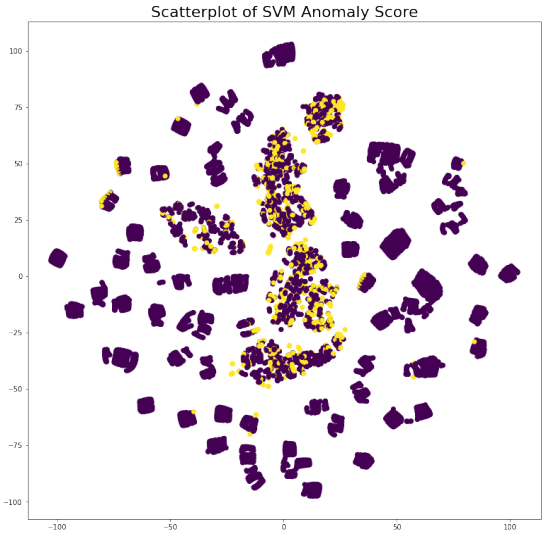
Subspace Outlier Detection

Общая идея - найти множество референсных точек, например, ближайших соседей. Используя подмножество признаков создать гиперплоскость вдоль которой лежит максимальное количество признаков. И затем посмотреть как далеко лежит от этой плоскости лежит наша точка.



One-class SVM detector

Основная идея – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве

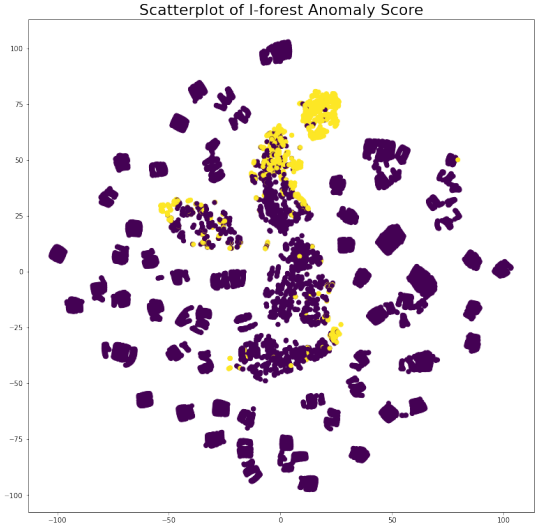


Higher School of Economics
Faculty of Computer science

Модели поиска аномалий

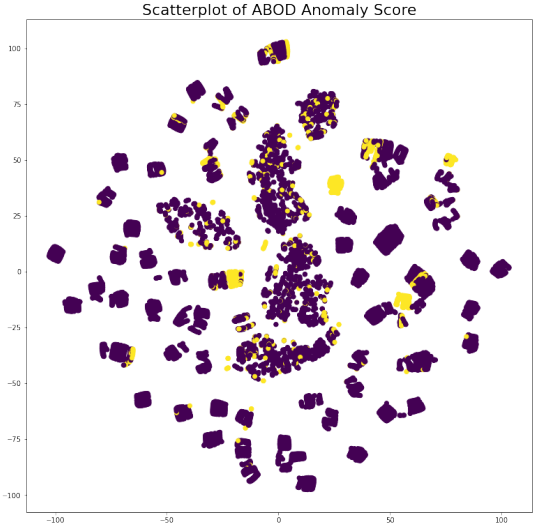
Isolation forest

Основная идея - пробуем "изолировать" наблюдение от всех остальных, и посмотреть насколько это легко можно сделать. Если слишком легко, то, скорее всего наблюдение лежит далеко и является выбросом. Если очень тяжело - скорее всего она похожа на многие другие точки и выбросом не является.



Angle-base Outlier Detection

Основная идея - Есть кластер точек. Если точки находятся внутри этого условного кластера, то угол между векторами связывающего данную точку попарно с другими постоянно меняется в зависимости от выбранных точке. Если же точка лежит вне кластера, то перебирая точки угол между векторами меняется незначительно.

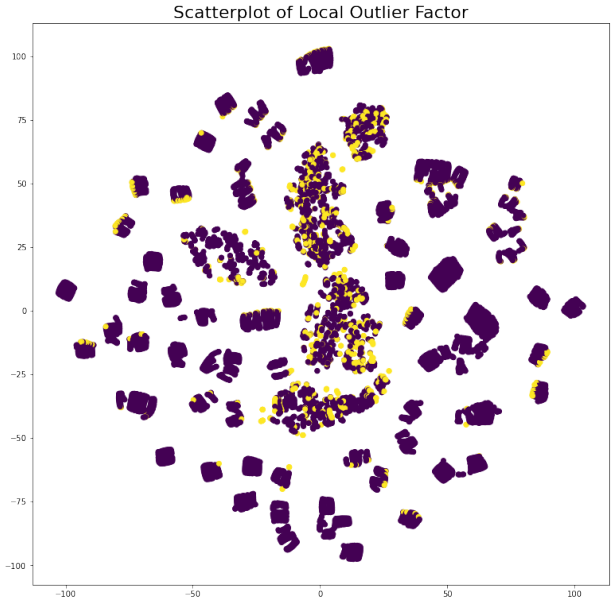


Higher School of Economics
Faculty of Computer science

Модели поиска аномалий

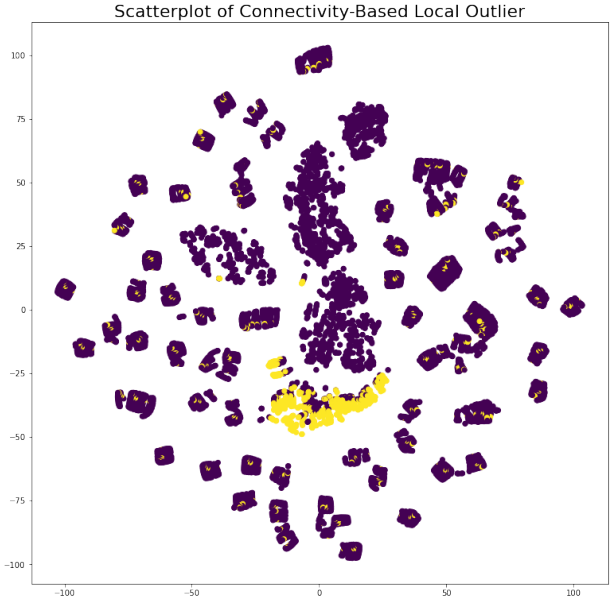
Local Outlier Factor

Основная идея – сравнение плотности данных вокруг рассматриваемой точки и вокруг ее соседей. Вокруг выброса гораздо меньшая плотность соседних точек.



Connectivity -Based Local Outlier

Основная идея – доработать LOF, чтобы учитывать не просто плотность точек, а количество граней графа последовательного доступа до ближайших точек.

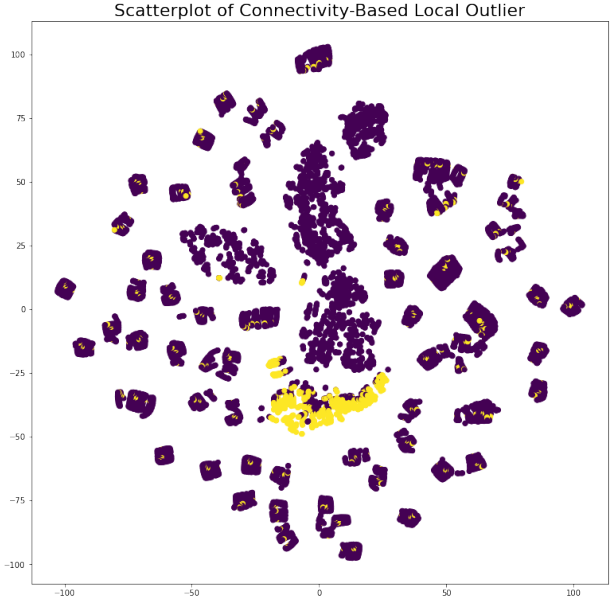


Higher School of Economics
Faculty of Computer science

Модели поиска аномалий

Clustering -Based Local Outlier

Основная идея - разбить данные на кластеры (на основании алгоритма К-средних), отранжировать на большие и малые кластеры, и для каждой точки из малого кластера потом посчитаем расстояние до ближайшего большого и на основании этого составим счет (anomaly score) и примем решение - аномалия точка или нет.



Histogram -based Outlier Detection

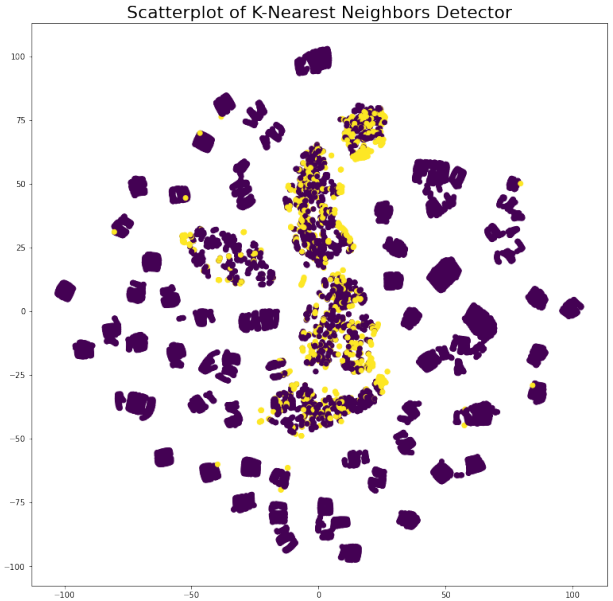
Основная идея – Для каждого измерения d построить одномерную гистограмму, где высота каждой ячейки отражает оценку плотности. Пронормировать, чтобы максимальная гистограмма была 1 и для каждой ячейки произвести расчет показателя аномалии на основе высоты шкалы.

Higher School of Economics
Faculty of Computer science

Модели поиска аномалий

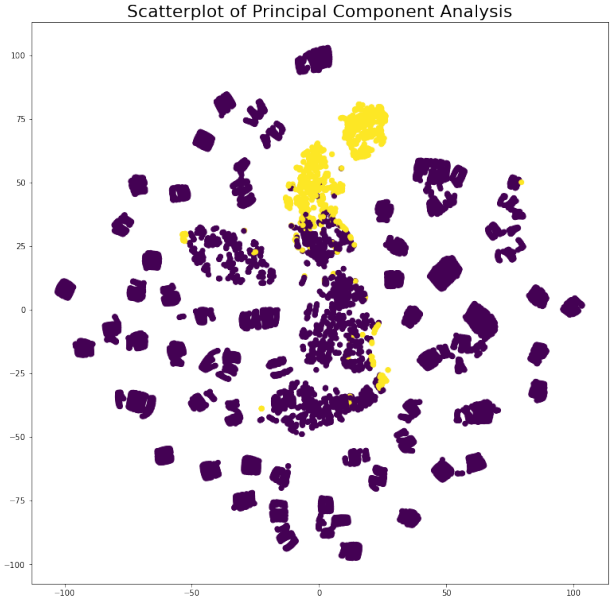
K-Nearest Neighbors Detector

Основная идея – для точки расстояние до K-соседей может рассматриваться как показатель аномалии



Principal Component Analysis

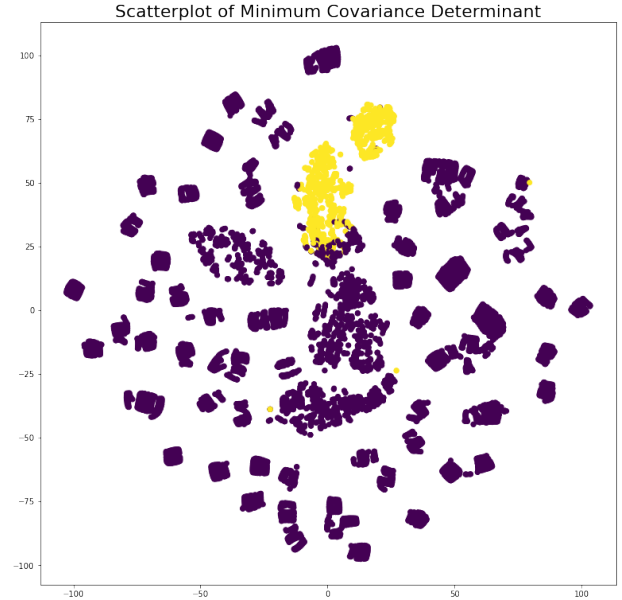
Основная идея – снижаем размерность N-признакового пространства чтобы удобно было разделить признаки гиперплоскостями. Затем как показатель аномалии посчитать расстояние от точки до гиперплоскости

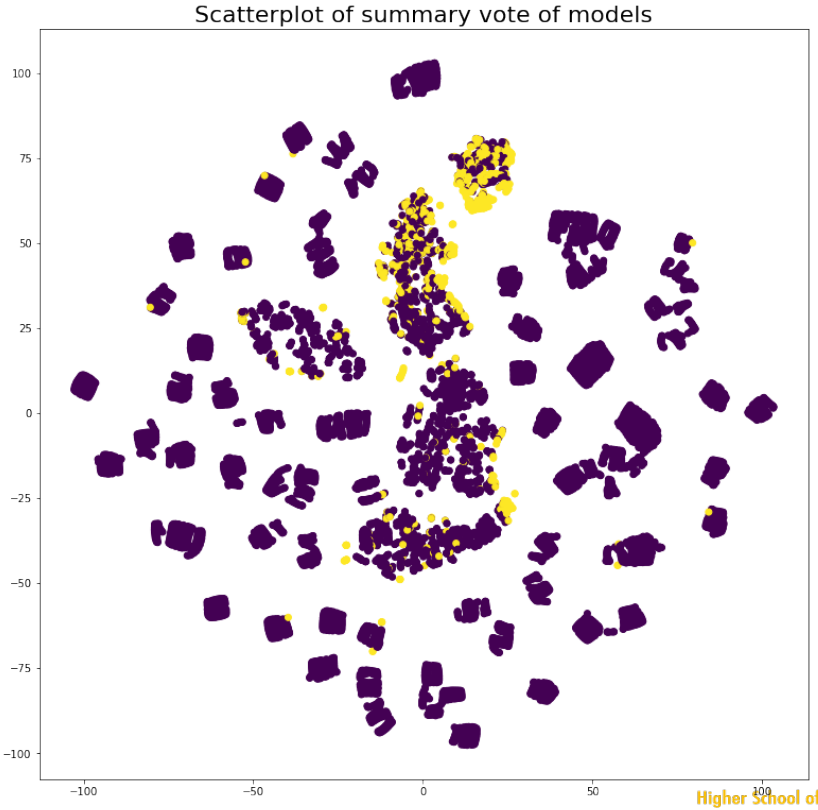


Модели поиска аномалий

Minimum Covariance Determinant

Основная идея – предполагаем что совокупность наших признаков подчинена многомерному распределению (нормальное) и определяем центр распределения и разброс, при помощи вектора средних и ковариционной матрицы. И определяем точки наиболее далеко расположенные от центра этого распределения.

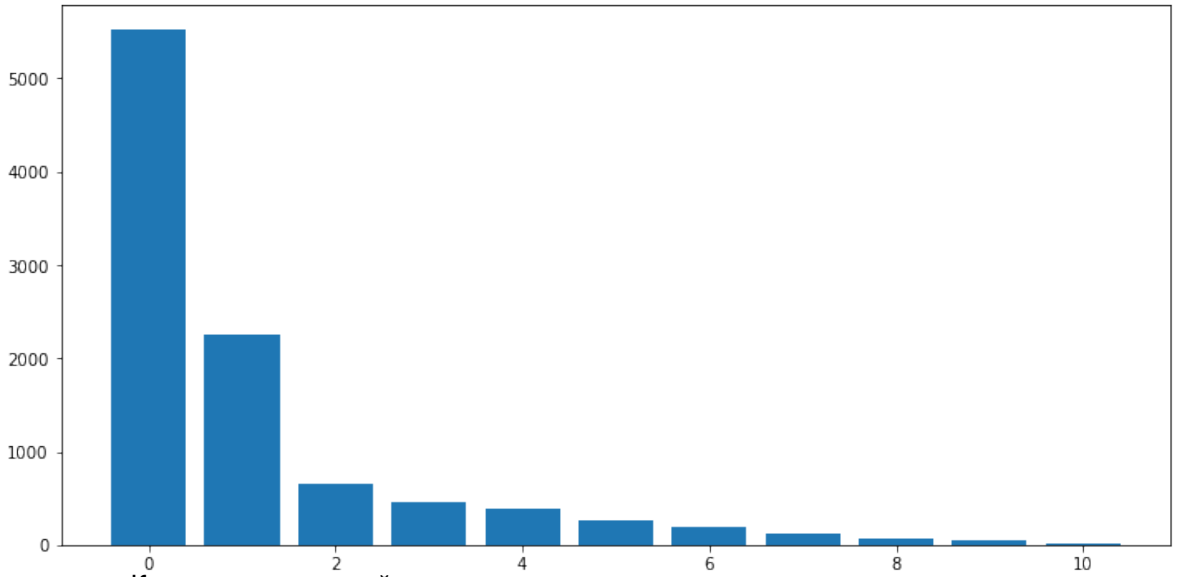




Higher School of Economics
Faculty of Computer science

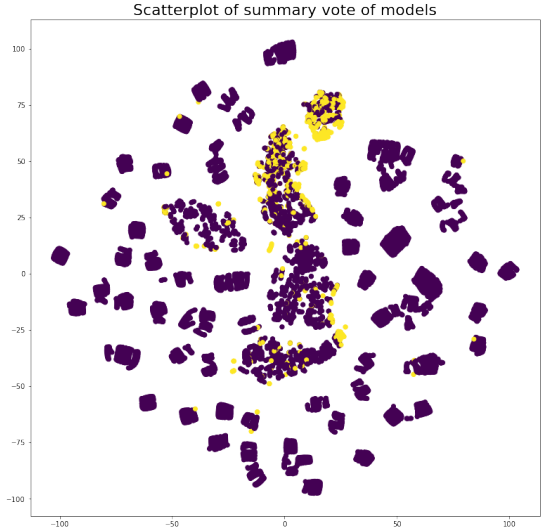
Ансамблирование методов

Кол-во точек

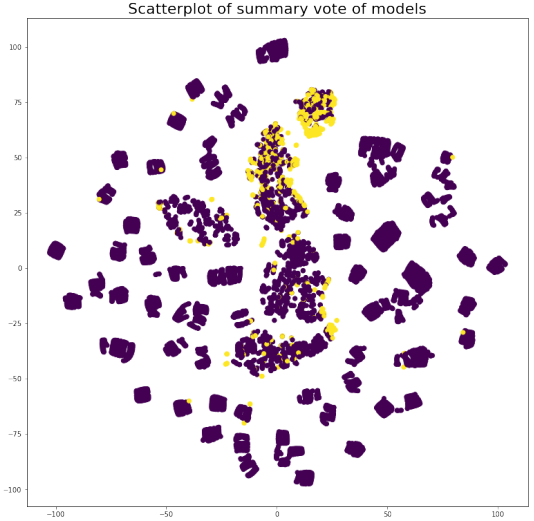


модели	точки
0	5521
1	2245
2	653
3	453
4	392
5	269
6	185
7	120
8	75
9	43
10	19
11	0

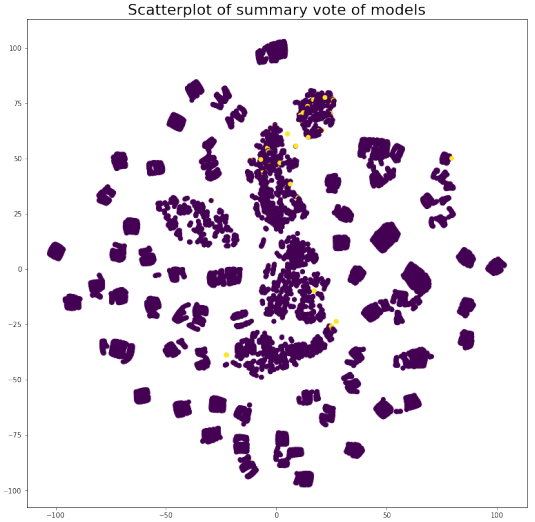
Диаграмма аномалий по итогам ансамблирования методов



N=5



N=8



N=10

N – количество моделей, проголосовавших за аномалию

Сравнение моделей поиска аномалии (id установки = 226000188)

model_name	accuracy_score	recall_score	f1_score	precision_score
knn_an	**0.942356**	**0.797468**	**0.663546**	**0.568136**
pca_an	**0.939950**	**0.780591**	**0.649503**	**0.556112**
histogram_an	**0.933935**	**0.738397**	**0.614394**	**0.526052**
iforest_an	0.932932	0.731364	0.608543	0.521042
svm_an	0.927519	0.693390	0.576946	0.493988
mcd_an	0.925915	0.682138	0.567583	0.485972
lof_an	0.915890	0.611814	0.509070	0.435872
sod_an	0.909273	0.565401	0.470451	0.402806
abod_an	0.876792	0.337553	0.280866	0.240481
cof_an	0.865564	0.258790	0.215331	0.184369
cluster_an	0.844712	0.092827	0.078525	0.068041

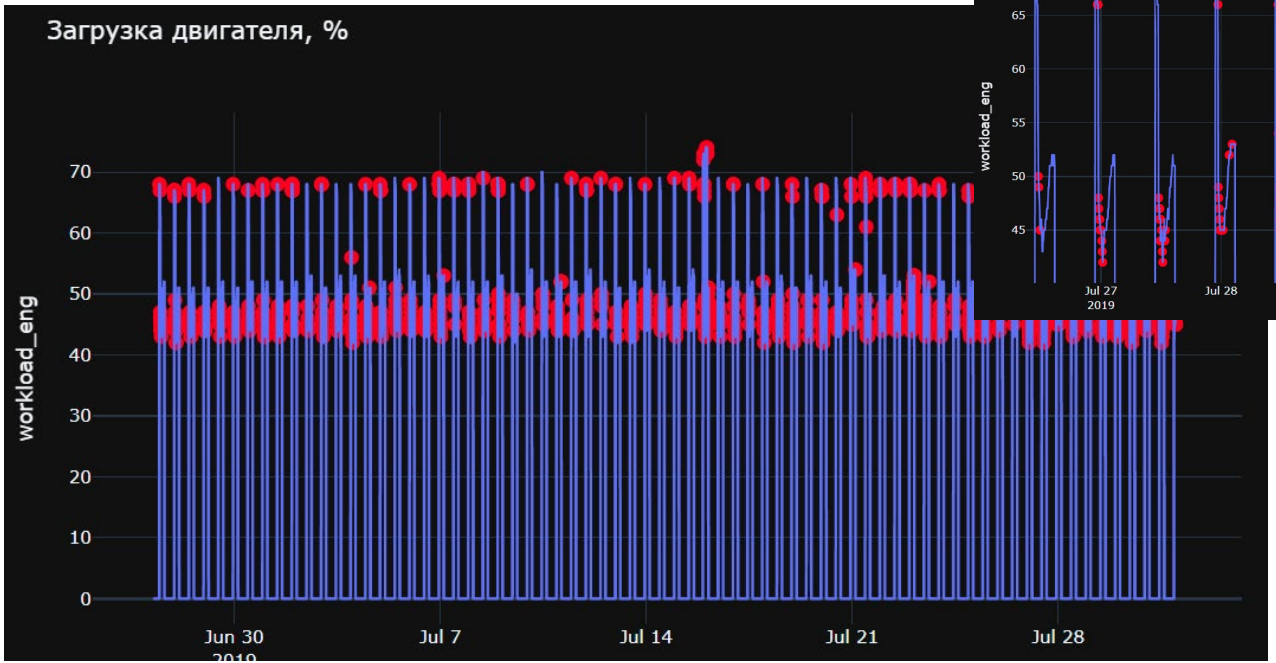
Метрики при сравнении предсказания модели и ансамблированного результата для **5** моделей – т.е. считаем что если за точку проголосовало 5 моделей, то это аномалия.

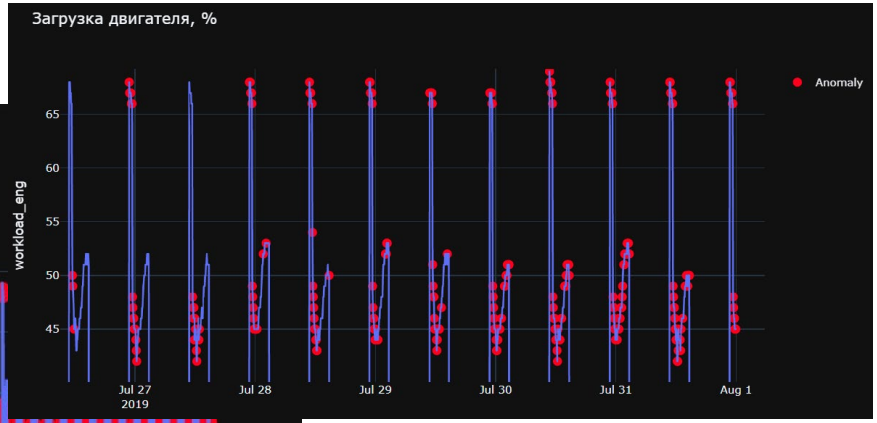
Сравнение моделей поиска аномалии (17 насосов)

model_name	accuracy_score	recall_score	f1_score	precision_score
knn_an	**0.940024**	**0.798447**	**0.627838**	**0.529481**
svm_an	**0.928465**	**0.706016**	**0.557811**	**0.471711**
pca_an	**0.926297**	**0.700766**	**0.548282**	**0.460870**
iforest_an	0.921623	0.662130	0.519861	0.437501
histogram_an	0.921262	0.664575	0.519285	0.435693
mcd_an	0.913973	0.602470	0.473293	0.399268
sod_an	0.907905	0.562947	0.439294	0.368943
lof_an	0.904274	0.525368	0.414058	0.350787
cof_an	0.892047	0.443305	0.343991	0.289667
abod_an	0.876727	0.295276	0.245014	0.213081
cluster_an	0.874128	0.288201	0.231884	0.197118

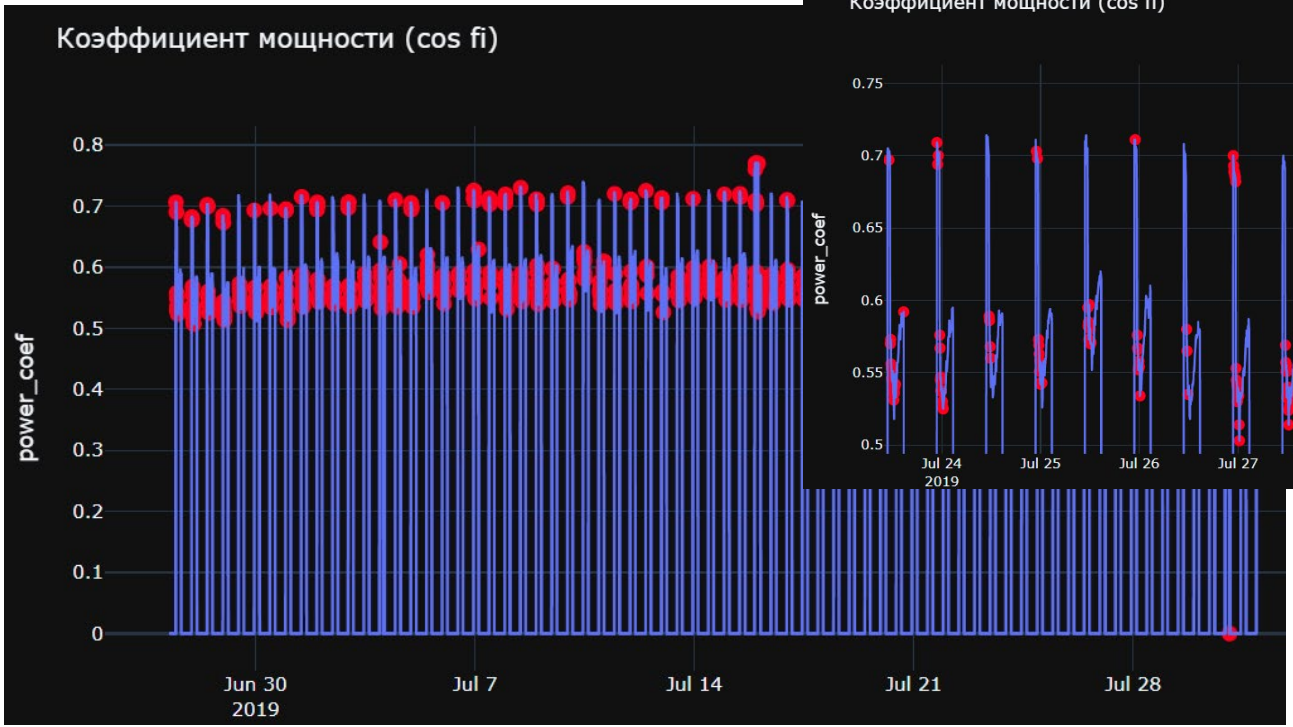
Метрики при сравнении предсказания модели и ансамблированного результата для **5** моделей – т.е. считаем что если за точку проголосовало 5 моделей, то это аномалия.

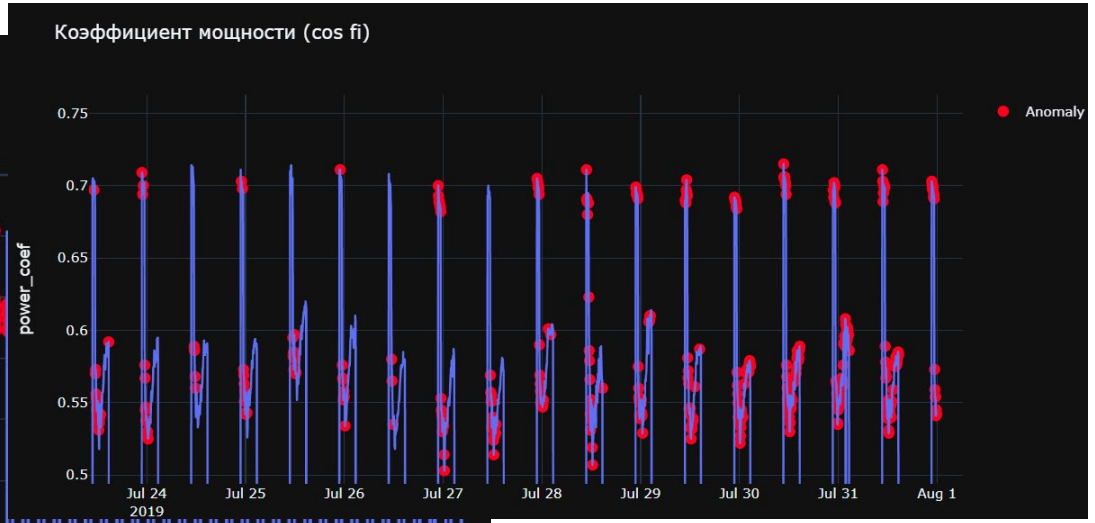
Модели поиска аномалий





Модели поиска аномалий





Higher School of Economics
Faculty of Computer science

Выводы

- **Был проведен сравнительный анализ моделей поиска аномалий и дана предварительная оценка качества моделей**

- **Финальная оценка будет возможна по итогам подтверждения техническим специалистом предполагаемых аномалий**

- **Лучшими моделями на представленном датасете предварительно показали себя модели:**

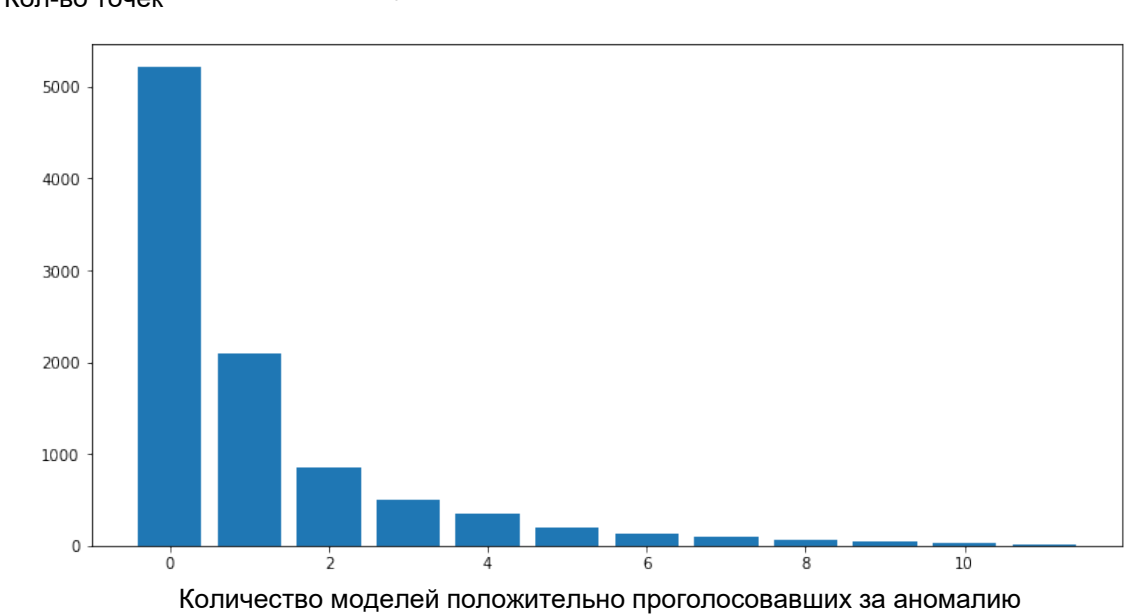
  Метод KNN (K-Nearest Neighbors Detector)
  SVM (One-class SVM detector)
  PCA (Principal Component Analysis)

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

Дополнительные материалы

Ансамблирование методов

Id установки = 226003080

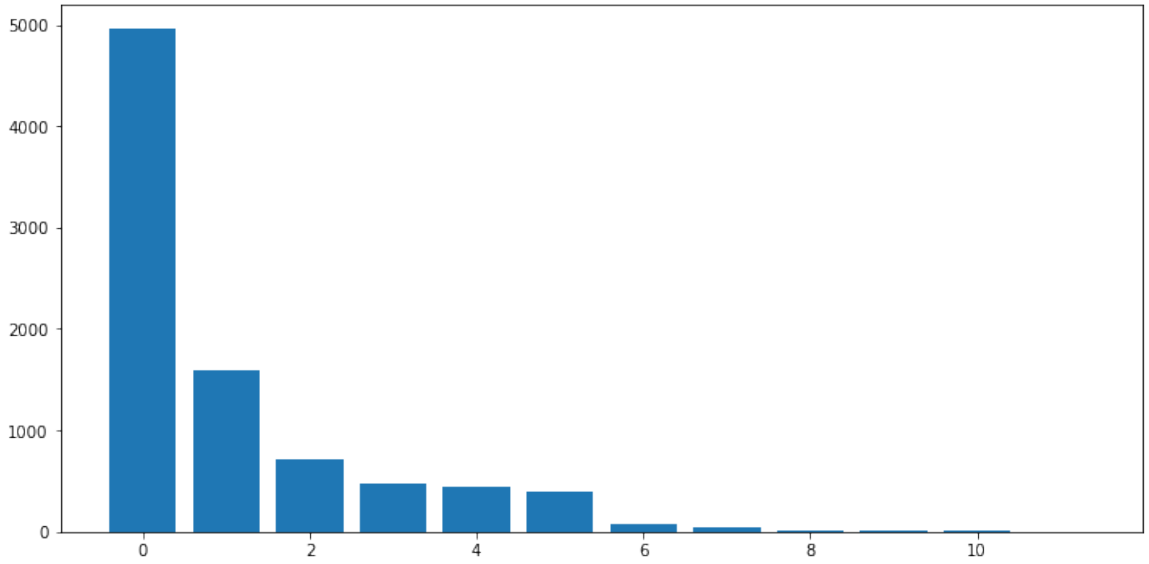


модели	точки
0	5214
1	2102
2	853
3	500
4	355
5	196
6	127
8	89
7	71
9	55
10	32
11	6

Ансамблирование методов

Кол-во точек

Id установки = 226002998



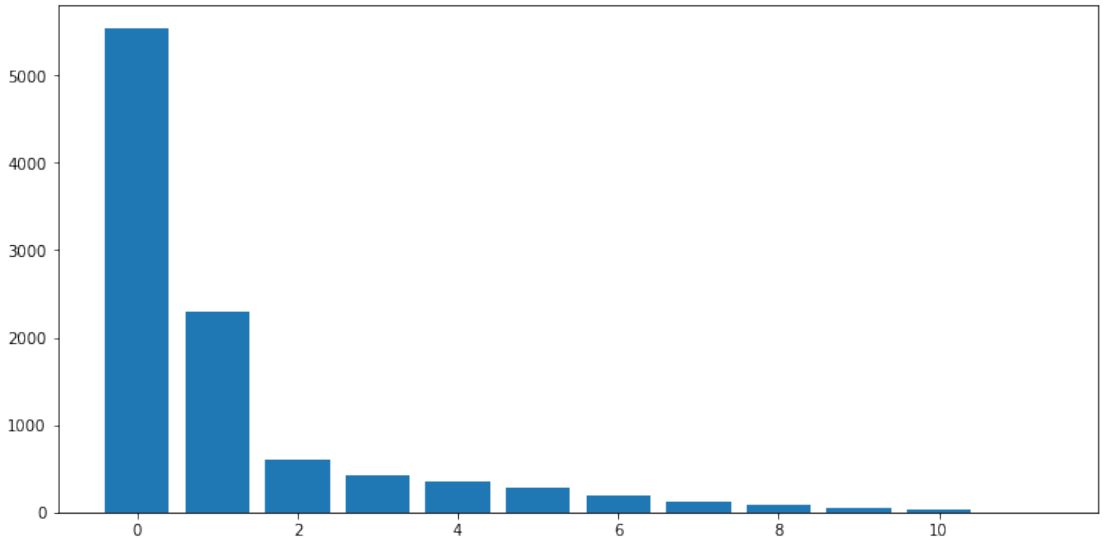
Количество моделей положительно проголосовавших за аномалию

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

Ансамблирование методов

Кол-во точек

Id установки = 226000297



Количество моделей положительно проголосовавших за аномалию

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

Сравнение моделей поиска аномалии (id установки = 226000188)

model_name	accuracy_score	recall_score	f1_score	precision_score
knn_an	**0.924311**	**0.972763**	**0.398406**	**0.250501**
svm_an	**0.923910**	**0.964981**	**0.395219**	**0.248497**
lof_an	**0.920702**	**0.902724**	**0.369721**	**0.232465**
pca_an	0.919699	0.883268	0.361753	0.227455
iforest_an	0.917694	0.844358	0.345817	0.217435
histogram_an	0.914085	0.774319	0.317131	0.199399
mcd_an	0.910276	0.700389	0.286853	0.180361
sod_an	0.908471	0.665370	0.272510	0.171343
abod_an	0.903058	0.560311	0.229482	0.144289
cof_an	0.897644	0.455253	0.186454	0.117234
cluster_an	0.883409	0.124514	0.052160	0.032990

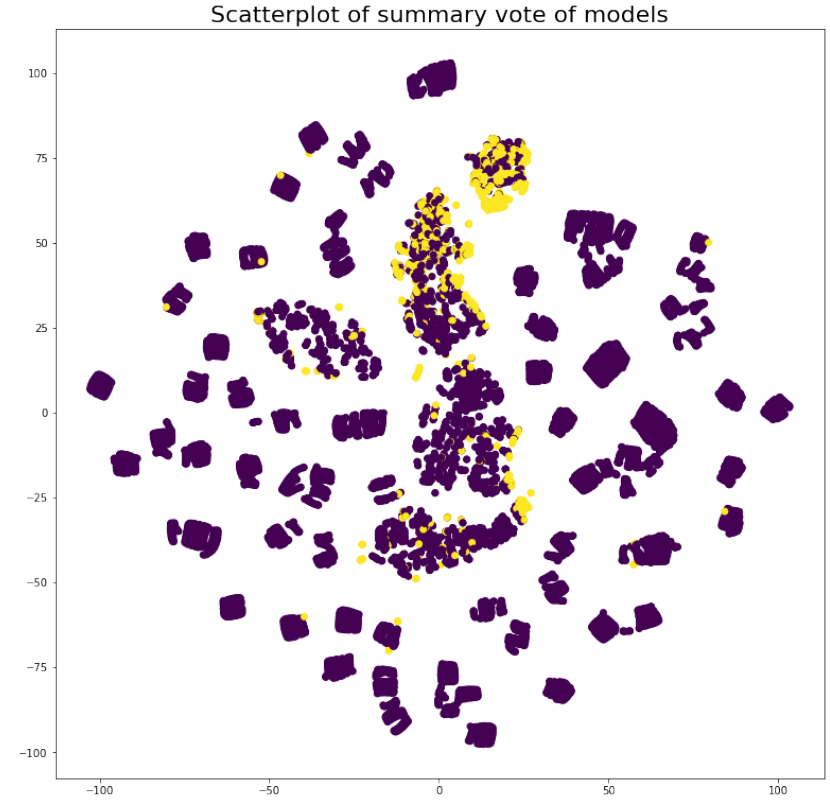
Метрики при сравнении предсказания модели и ансамблированного результата для **8** моделей – т.е. считаем что если за точку проголосовало 8 моделей, то это аномалия.

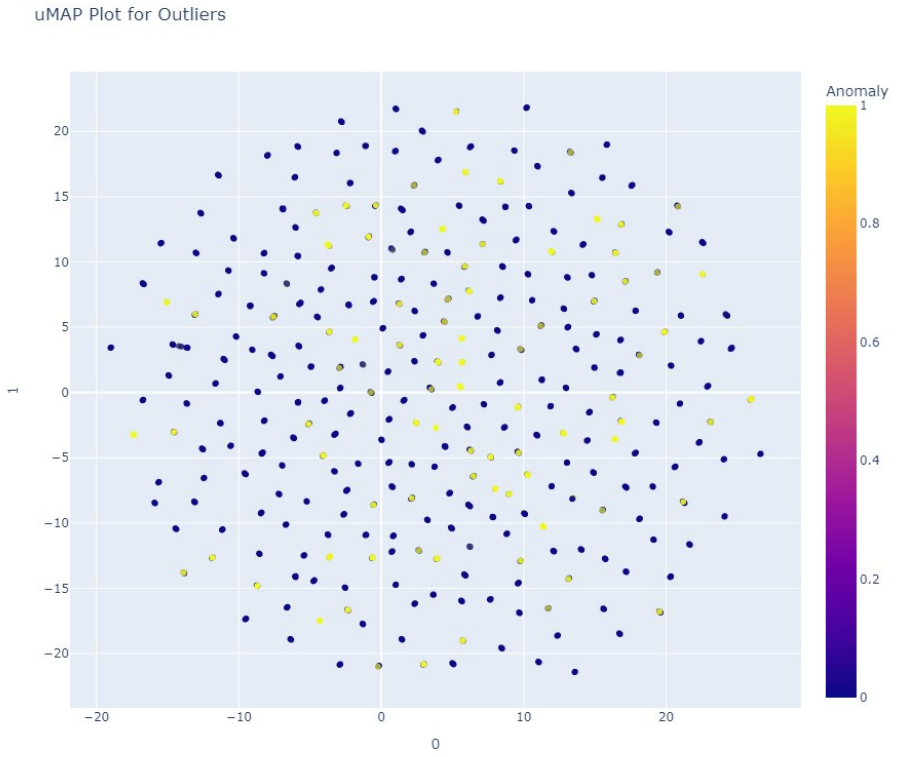
Сравнение моделей поиска аномалии (id установки = 226000188)

model_name	accuracy_score	recall_score	f1_score	precision_score
iforest_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
abod_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
cof_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
knn_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
lof_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
svm_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
pca_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
sod_an	**0.901855**	**1.000000**	**0.037365**	**0.019038**
histogram_an	0.901654	0.947368	0.035398	0.018036
mcd_an	0.901454	0.894737	0.033432	0.017034
cluster_an	0.901454	0.157895	0.006067	0.003093

Метрики при сравнении предсказания модели и ансамблированного результата для **10** моделей – т.е. считаем что если за точку проголосовало 10 моделей, то это аномалия.

Диаграмма аномалий по итогам голосования моделей n=5 (TSNE, uMap)





Higher School of Economics
Faculty of Computer science

Сравнение моделей поиска аномалии (17 насосов)

model_name	accuracy_score	recall_score	f1_score	precision_score
knn_an	**0.912692**	**0.980204**	**0.218371**	**0.129066**
svm_an	**0.912690**	**0.974932**	**0.218673**	**0.129295**
pca_an	**0.911772**	**0.963185**	**0.211590**	**0.124702**
iforest_an	0.911136	0.930469	0.205980	0.121523
histogram_an	0.910416	0.929303	0.200275	0.117835
lof_an	0.908975	0.850900	0.187241	0.110723
mcd_an	0.908943	0.797534	0.187258	0.110562
sod_an	0.906140	0.758163	0.163953	0.096557
cof_an	0.905613	0.710605	0.158087	0.093915
abod_an	0.901068	0.447457	0.116536	0.071184
cluster_an	0.896058	0.260001	0.070137	0.041713

Метрики при сравнении предсказания модели и ансамблированного результата для **8** моделей – т.е. считаем что если за точку проголосовало 5 моделей, то это аномалия.

Сравнение моделей поиска аномалии (17 насосов)

model_name	accuracy_score	recall_score	f1_score	precision_score
knn_an	**0.902649**	**0.823529**	**0.048975**	**0.026197**
svm_an	**0.902590**	**0.823529**	**0.048952**	**0.026184**
pca_an	**0.902578**	**0.819853**	**0.048836**	**0.026125**
iforest_an	0.902566	0.807315	0.048721	0.026066
histogram_an	0.902567	0.816139	0.048510	0.025956
lof_an	0.902505	0.816563	0.048154	0.025759
mcd_an	0.902463	0.789389	0.047719	0.025548
abod_an	0.902359	0.687674	0.046643	0.025005
cof_an	0.902174	0.770144	0.044961	0.024104
sod_an	0.902046	0.780168	0.043892	0.023466
cluster_an	0.900372	0.374323	0.019348	0.010184

Метрики при сравнении предсказания модели и ансамблированного результата для **10** моделей – т.е. считаем что если за точку проголосовало 5 моделей, то это аномалия.

Центр непрерывного образования

Специалист по Data Science

программа профессиональной переподготовки