ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК XXXXX

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ФИНАНСОВЫХ ДАННЫХ

Выполнили:

студент группы МОВС 2021
Кулакин Сергей Алексеевич

_____ _____
(подпись) (дата)

студентка группы МОВС 2021
Журович Янина Александровна

_____ _____
(подпись) (дата)

Руководитель:

Преподавательница
Факультета компьютерных наук НИУ ВШЭ
Максимовская Анастасия Максимовна

_____ _____
(подпись) (дата)

Москва, 2023

Содержание

Аннотация

1. Введение

Начать стоит с того, что под аномалиями понимаются необычные части данных, выбросы, шумы, отклонения или исключения. В анализе данных существуют два направления, которые занимаются поиском аномалий: *детектирование выбросов* (outlier detection) и *поиск новизны* (novelty detection). Как и выброс «новый объект» — это объект, который отличается по своим свойствам от объектов (обучающей) выборки, но в отличие от выброса, его в самой выборке пока нет (он появится через некоторое время, и задача заключается в том, чтобы обнаружить его при появлении). Например, при анализе замеров температуры отбрасываются аномально большие или наоборот маленькие значения, то есть происходит борьба с выбросами. А если создаётся алгоритм, который для каждого нового замера оценивает, насколько он похож на прошлые, и выбрасывает аномальные, то происходит «борьба с новизной».

Выбросы могут являться следствием нескольких факторов, к примеру:

- ошибок в данных (неточности измерения, округления, неверной записи и т.п.)
- наличия шумовых объектов (неверно классифицированных объектов)
- присутствия объектов «других» выборок (например, показаниями сломавшегося датчика).

Новизна в свою очередь, как правило, появляется в результате принципиально нового поведения объекта. Предположим, что объекты – это описания работы системы, тогда после проникновения в неё вируса объекты становятся «новизной». Или другим примером могут являться описания работы двигателя после поломки.

Как уже понятно из описания выше, процесс выявления аномалий заключается в распознавании редких данных, событий или наблюдений в ходе data mining, которые вызывают подозрения ввиду существенного отличия от большей части данных. Этот процесс зачастую необходим, так как реальные данные, собранные для анализа или решения каких-то определённых задач, далеки от идеальных. Как правило, такие аномалии характеризуют некоторый вид проблемы. К примеру, это могут быть ошибки в тексте, медицинские затруднения в процессе мониторинга здоровья, какие-либо структурные дефекты и неисправности, нарушения в экологической сфере или мошенничество в банке — случай, который как раз и будет рассматриваться далее.

Постановка проблемы и её актуальность

Для тех, кто работал с реальными данными, не является секретом тот факт, что для построения хороших моделей данные требуется тщательно предобработать. В том числе избавиться от тех самых аномалий, объяснить их.

В таком случае, актуальность проблемы поиска различных аномалий в данных и следующие за этим закономерные действия очевидно необходимы и могут быть применимы в задачах любой сложности и направленности.

В рамках представленного исследования предположим, что выдуманный банк хотел бы по транзакциям клиентов определять, является ли конкретная транзакция или их группа подозрительной(-ыми).

Цель работы и задачи

В настоящей работе *основной уклон* будет сделан на разборе существующих способов и методов борьбы с аномалиями: их поиском и дальнейшим преобразованием данных, а также применении методов машинного и глубинного обучений для решения поставленной задачи.

Объектом исследования является выбранные наборы данных с информацией о банковских транзакциях, включающих в себя мошеннические операции. *Предметом* исследования можно считать обнаруженные в данных аномалии, которые в процессе анализа подверглись различным изменениям и объяснениям. *Теоретическую основу* данной работы составляют труды отечественных и зарубежных исследователей, которые изучали данную проблему в разных случаях и перспективы её развития. *Методологическую базу* составляют такие методы исследования, как системно-аналитический, наблюдение, сопоставление, моделирование, проведение экспериментов и тестирования. *Эмпирической базой* являются информационные, статистические базы и собрания, а также нормативные документы и датасеты.

В этом случае *основные задачи* могут быть сформулированы предварительно следующим образом:

– Сбор и/или поиск подходящего датасета.

Для данного исследования были выбраны два датасета:

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download

Популярный и часто используемый набор данных, являющийся достаточно обкатанным и представляющий собой информацию о транзакциях, совершённых по кредитным картам держателями из Европы в сентябре 2013 года. К его особенностям можно отнести ощутимую несбалансированность: так, на положительный класс (мошенничество) приходится 0.172% всех транзакций. Помимо этого, в нём содержатся только числовые входные данные, являющиеся результатом преобразований PCA, так как из соображений конфиденциальности нет возможности предоставления исходных функций и дополнительной справочной информации (далее – CCF).

[2] https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?resource=download

Набор данных 2022 года, представляем мошенничество с банковскими счетами (BAF), первый общедоступный, крупномасштабный, реалистичный набор табличных данных с сохранением конфиденциальности. Набор создан путем применения современных методов генерации табличных данных (далее – BAF).

– EDA подозрительных транзакций (попытка найти ответ на вопрос, в проведение каких именно транзакций банку стоит дополнительно вмешаться):

- Рассчитать основные статистики для вещественных и категориальных признаков;
- Выяснить, есть ли пропущенные значения и сколько; какова их доля от общего числа объектов для каждого конкретного признака (и, в целом, насколько выбранный датасет “разрежен” в контексте пропущенных значений). Затем выдвинуть и проверить гипотезу о том, что могут означать пропущенные значения. При этом осмысленно и аргументированно обработать пропуски;
- Есть ли ошибочные (например, отрицательный возраст или пятиметровый рост человека; 3-й класс в задаче бинарной классификации) значения (признаки, целевая переменная) в данных. Если такие имеются, обработать их;
- Определить, есть ли выбросы в данных? По возможности обработать их.

– Визуализация в виде графиков и гистограмм;

– Применение и решение ML-задач:

- построение классификатора аномальных транзакций;
- балансировка данных;

- статистические подходы к поиску аномалий;
- подход без учителя (кластеризация, ML-методы поиска аномалий);
- объединение двух подходов.

– Улучшение прогноза:

- объединение подходов обучения с и без учителя в semi-supervised способ;

– Применение и решение DL-задач:

- GAN сеть для генерации неподозрительных транзакций (прогон транзакции через дискриминатор).

Методология и теоретическая ценность

Обратившись к математическому аппарату с учётом уже применяемых методов на практике, решено было использовать на данном этапе следующее:

– EDA датасетов;
– CatBoost;
– LGBM;
– Hist Gradient Boosting;
– Logistic Regression;
– Decision Tree;
– Random Forest;
– Multi-layer Perceptron classifier (MLP);
– Stacking Classifier (Decision Tree, Random Forest, Hist Gradient Boosting);
– Одноклассовый SVM;
– Isolation Forest

Практическая значимость

Как было упомянуто ранее, итоги представленного исследования могут быть использованы в схожих задачах по поиску мошеннических транзакций банками и иными финансовыми институтами.

Хотелось бы также отметить, что данное исследование и пример применения различного набора моделей не является истинно верным для любых датасетов или задач, а подходящим решением, на взгляд исследователей, в данной ситуации и схожих с ними.

2. Обзор литературы

В настоящей главе будут рассмотрены некоторые более классические подходы, современные теории и методы, данные предшествующих различных исследований на выбранную тематику, концептуальная схема исследования, а также частичное обоснование теоретических рамок данной работы.

Стоит начать с того, что в уже далеком 1986 году американская исследовательница и ученый в области информационной кибербезопасности Дороти Деннинг одной из первых предложила выявление аномалий для систем обнаружения вторжений [4]. Стоит отметить, что выявление аномалий для подобных систем по обнаружению вторжений обычно выполняется с заданием порога и статистики, однако может быть сделано и с помощью мягких вычислений и индуктивного обучения [11]. При этом, типы статистики, предлагавшиеся в то время, включали профайлы пользователей, рабочих станций, сетей, удалённых узлов, групп пользователей или программ, основанных на частотах, средних и дисперсиях. А эквивалентом выявления аномалий в обнаружении вторжений является обнаружение злонамеренного использования. Тем временем, параллельно и также в дальнейшем детектирование аномалий в данных применялось практически повсеместно и при решении большого количества задач различного спектра. Как было сказано ранее, это могли быть задачи от сейсмологии или поиска нестандартных игроков на бирже (инсайдеров) до обнаружения подозрительных банковских операций (credit-card fraud) или неполадок в механизмах по показаниям датчиков.

Далее стоило бы остановиться на кратком рассмотрении общих методик обнаружения аномальных показателей. Так, к примеру, это могут быть:

– Статистические тесты (как правило, применяют для отдельных признаков и отлавливают экстремальные значения (extreme-value analysis). Для этого используют, например, z-value или kurtosis measure);

– Модельные тесты (строим модель, которая описывает данные, а точки, которые сильно отклоняются от модели (на которых модель сильно ошибается) и будут являться аномалиями; также при выборе модели можно учесть природу задачи, функционал качества и т.п. Такие методы хорошо работают для определения новизны, но хуже справляются при поиске выбросов, потому что при настройке модели используются данные, в которых уже есть выбросы (и она под них «затачивается»);

– Итерационные методы (состоят из итераций, на каждой из которых удаляется группа «особо подозрительных объектов»);

– Метрические методы (судя по всему, самые популярные методы среди исследователей: в них постулируется существование некоторой метрики в пространстве объектов, которая и помогает найти аномалии. Интуитивно понятно, что у выброса мало соседей, а у типичной точки много. Поэтому хорошей мерой аномальности может служить, например «расстояние до k-го соседа»);

– Методы подмены задачи (при возникновении и необходимости решения новой задачи можно применить ориентированные на уже известные задачи методы);

– Методы машинного обучения (можно решать задачу не как, например, классификацию или кластеризацию, а использовать, допустим, метод опорных векторов, изолирующий лес, эллипсоидальную аппроксимацию данных и т. д.);

– Ансамбли алгоритмов (есть мнение, что использование сразу несколько методов практически никогда не является плохой идеей, поэтому часто в подобного рода задачах используются некоторые виды бэггинга и т. д.).

Заканчивая краткий экскурс в историю, следовало бы перейти к обзору новейших методов для решения выбранной задачи. Так, например, в статье [2] предлагается новый подход, называемый Fuzzy Isolation Forest, который адаптирует метод Isolation Forest к нечетким данным для обработки их «неопределенности». В таком случае используется стратегия нечеткого разбиения, основанная на методе alpha-cut (альфа-вырезания), при котором экземпляры данных разделяются в узлах дерева путем сравнения их значений с порогом альфа-вырезания. Fuzzy Isolation Forest так же эффективен, как и Isolation Forest, в обнаружении аномалий в нечетких данных без значительного увеличения времени обработки.

Помимо этого, существует подход гибридного машинного обучения для обнаружения аномалий. Исследование [8] направлено на повышение эффективности обнаружения аномалий путем разработки двух вариантов гибридных моделей, сочетающих методы машинного обучения с учителем и без учителя. Модели с учителем не могут обнаруживать новые типы аномалий. В первом варианте за моделью с учителем, которая обнаруживает нормальные образцы, следует модель обучения без учителя для выявления аномалий. Во втором случае гибридная модель включает в себя модель обучения без учителя, которая обнаруживает аномалию, за которой следует модели с учителем для подтверждения этой аномалии.

Кроме того, можно утверждать, что в последние годы Isolation forest (iForest) стал фактически самым популярным детектором аномалий благодаря его общей эффективности в различных тестах и хорошей масштабируемости. Введено несколько расширений iForest, но в основном они по-прежнему используют неглубокое линейное разделение данных, ограничивая их возможности в выделении истинных аномалий. А вот в статье [5] предлагается как раз лес глубокой изоляции. А именно, вводится новая схема представления, в которой используются случайно инициализированные нейронные сети для отображения исходных данных в ансамбли случайных представлений, где впоследствии применяются случайные параллельные оси разделены для выполнения разделения данных. Обширные эксперименты показывают, что модель обеспечивает значительное улучшение по сравнению с современными методами, основанными на изоляции, и глубокими детекторами на наборах данных в виде таблиц, графиков и временных рядов. Модель также наследует желаемую масштабируемость от iForest.

Теперь рассмотрим исследовательскую работу [7], в которой предлагается новый метод, называемый автоэнкодером с адаптивной функцией потерь (AEAL), для повышения точности обнаружения известных аномалий, обеспечивающий при этом согласованные оценки аномалий до и после обновлений модели. AEAL — это метод на основе автоэнкодера для обучения моделей обнаружения аномалий путем динамической настройки баланса между минимизацией ошибок реконструкции для данных об аномалиях и максимизацией ошибок реконструкции для данных об аномалиях.

В данном разделе предложена работа [6], в которой применяется критерий эквивалентности для алгоритмов обнаружения аномалий, измеряющий, в какой степени два алгоритма обнаружения аномалий обнаруживают одинаковые виды аномалий. 1) Излагается набор желаемых свойств, которыми должен обладать такой критерий эквивалентности и почему. 2) Предлагается, критерий эквивалентности Гаусса (GEC) в качестве критерия эквивалентности и математически показано, что он обладает желаемыми свойствами, упомянутыми ранее. 3) Эмпирически подтверждены эти свойства с использованием смоделированных и реальных наборов данных. 4) Для реального набора данных показано, как GEC может дать представление об алгоритмах обнаружения аномалий, а также о наборе данных.

В дополнение рассмотрим статью [10] с всесторонним обзором алгоритмов обнаружения аномалий. В нём достаточно полно и организованно представлены алгоритмы обнаружения аномалий. 1) Обзор начинается с определения аномалии, основных элементов обнаружения аномалии, различных типов аномалий, разных областей применения и мер оценки. 2) Алгоритмы обнаружения аномалий подразделяются на семь категорий в зависимости от их рабочих

механизмов, которые включают в себя в общей сложности 52 алгоритма. 3) Категории представляют собой алгоритмы обнаружения аномалий, основанные на: статистике, плотности, расстоянии, кластеризации, изоляции, ансамбле, подпространстве. 4) Для каждой категории в статье предоставлена временная сложность каждого алгоритма, а также их общие преимущества и недостатки.

Ещё один пример работы с аномалиями в данных представлен в публикации [9], где фигурирует метод словаря шаблонов. Он основан на сжатии метода обнаружения аномалий для временных рядов и данных последовательностей с использованием словаря шаблонов. 1) Предлагаемый метод способен изучать сложные закономерности в последовательности обучающих данных, используя эти изученные закономерности для обнаружения потенциально аномальных закономерностей в последовательности тестовых данных. 2) Предлагаемый метод словаря шаблонов использует меру сложности тестовой последовательности в качестве оценки аномалии, которую можно использовать для автономного обнаружения аномалии. 3) Также показано, что в сочетании с универсальным исходным кодером предлагаемый словарь шаблонов дает мощный детектор атипичности, который в равной степени применим для обнаружения аномалий.

Более того, существует такое понятие, как надежное обнаружение аномалий в критической инфраструктуре, описанное в [1]. О чём же здесь идёт речь? Важнейшие инфраструктуры (КИ), такие как водоочистные сооружения, электрические сети и телекоммуникационные сети, имеют решающее значение для повседневной деятельности и благополучия нашего общества. Нарушение таких КИ имело бы катастрофические последствия для общественной безопасности и национальной экономики. Следовательно, эти инфраструктуры стали основными мишенями для кибератак. Защита от таких атак часто зависит от арсенала инструментов киберзащиты, включая системы обнаружения аномалий (ADS) на основе машинного обучения (ML). Эти системы обнаружения используют модели ML для изучения профиля нормального поведения CI и классификации отклонений, выходящих далеко за рамки нормального профиля, как аномалий. Однако методы ML уязвимы как для враждебных, так и для непротиворечивых входных возмущений. Враждебные возмущения — это незаметные шумы, добавляемые злоумышленником к входным данным для обхода механизма классификации. Непротиворечивые возмущения могут быть нормальной эволюцией поведения в результате изменений в моделях использования или других характеристиках и зашумленных данных от нормально деградирующих устройств, генерируют высокий процент ложных срабатываний. Сначала изучается проблема уязвимости ADS на основе ML к непротиворечивым возмущениям, что вызывает высокий уровень ложных срабатываний. Для решения этой

проблемы предлагается ADS под названием DAICS, на основе широкой и глубокой модели обучения, которая одновременно адаптируется к развивающейся нормальности и устойчива к зашумленным данным, обычно поступающим из системы. DAICS адаптирует предварительно обученную модель к новой нормальности с помощью небольшого количества выборок данных и нескольких обновлений градиента на основе отзывов оператора о ложных тревогах. DAICS оценивается на двух наборах данных, собранных на реальных испытательных стендах Industrial Control System (ICS). Результаты показывают, что процесс адаптации проходит быстро и что DAICS обладает повышенной надежностью по сравнению с современными подходами. Далее была исследована проблема ложноположительных сигналов тревоги в iv ADS. Чтобы решить её, предлагается расширение DAICS, называемое структурой SiFA. Такая конструкция собирает буфер исторических ложных тревог и подавляет каждую новую тревогу, похожую на эти ложные тревоги. Предлагаемая структура оценивается с использованием набора данных, собранных на реальном испытательном стенде ICS. Результаты оценки показывают, что SiFA способна снизить частоту ложных срабатываний DAICS более чем на 80%. В работе также исследуется проблема сетевых ADS на основе ML, которые уязвимы для враждебных возмущений. В случае сетевых ADS, злоумышленники могут использовать свои знания логики обнаружения аномалий для создания вредоносного трафика, который остается необнаруженным. Одним из способов решения этой проблемы является использование состязательного обучения, в котором обучающая выборка дополняется состязательно возмущенными образцами. В этой диссертации представлен подход к состязательному обучению, называемый GADoT, который использует генеративно-состязательную сеть (GAN) для создания состязательных образцов для обучения. GADoT проверяется в сценарии, когда ADS обнаруживает распределенные атаки типа «отказ в обслуживании» (DDoS), объем и сложность которых увеличиваются. Для практической оценки сетевой трафик DDoS был изменен для создания двух наборов данных при полном сохранении семантики атаки. Результаты показывают, что злоумышленники могут использовать свой опыт в предметной области для создания состязательных атак, не требуя знания базовой модели обнаружения. Затем демонстрируется то, что состязательное обучение с использованием GADoT делает модели машинного обучения более устойчивыми к состязательным возмущениям. Однако оценка устойчивости к состязательным действиям часто подвержена ошибкам, что приводит к переоценке устойчивости. Таким образом, исследуется проблема переоценки надежности сетевых ADS и предлагается состязательная атака, называемая UPAS, для оценки надежности таких ADS. Атака UPAS нарушает время между приходами пакетов, вводя случайную временную задержку перед пакетами от злоумышленника. Она проверяется путем нарушения вредоносного сетевого трафика в наборе данных с несколькими атаками и используется для оценки надежности двух надежных ADS, которые основаны на шумоподавляющем автоэнкодере

и ML-модели, обученной со стороны злоумышленников. Результаты показывают, что надежность обеих ADS завышена и что необходима стандартизированная оценка надежности.

И наконец рассмотрим статью [3] о разделе, с которого было начато повествование. Благодаря постоянному развитию таких технологий, как интернет вещей (IoT) и облачные вычисления, датчики собирают и хранят большие объемы сенсорных данных, обеспечивая запись и восприятие окружающей среды в режиме реального времени. Из-за открытых характеристик WSN риски безопасности при передаче информации являются заметными, и сетевая атака или вторжение, вероятно, произойдет. Таким образом, эффективное обнаружение аномалий жизненно важно для систем IoT, чтобы обеспечить безопасность системы. Оригинальный алгоритм Isolation Forest представляет собой алгоритм обнаружения аномалий с линейной временной сложностью и имеет лучший эффект обнаружения на воспринимаемых данных. Однако, есть также недостатки, такие как сильная случайность, низкая производительность обобщения и недостаточная стабильность. В этой статье предлагается метод обнаружения аномалий данных под названием BS-iForest (iForest с коробчатой диаграммой) для беспроводных сенсорных сетей, основанный на варианте Isolation Forest для решения проблем. Такой метод сначала использует суб-набор данных, отфильтрованный по блочной диаграмме, для обучения и построения деревьев. Затем в обучающей выборке выбираются изолирующие деревья с более высокой точностью для формирования базового детектора аномалий леса. Затем детектор аномалий базового леса использует обнаружение аномалий для оценки выбросов данных в течение следующего периода. Эти эксперименты проводились на наборах данных, собранных с датчиков, развернутых в центре обработки данных университета, и на наборе данных груди Висконсина (BreastW), демонстрируя производительность варианта алгоритма Isolation Forest. По сравнению с традиционным изолированным лесом площадь под кривой (AUC) увеличилась на 1.5% и 7.7%, и это подтвердило то, что предложенный метод превосходит стандартный алгоритм Isolation Forest с двумя выбранными в данном случае наборами данных.

3. Разведочный анализ данных и подготовка данных

В ходе проведения разведочного анализа были выявлены следующие особенности в наборах данных.

1) В первом наборе (CCF) из 284 807 транзакций всего 492 мошеннических. Все признаки (31 переменная) в наборе числовые и являются результатом PCA разложения от реальных данных. Пропуски в данных отсутствуют. Из анализа графиков распределений переменных следует, что дополнительная предобработка данных не требуется.

Для проведения моделирования на данном наборе данных будем разделять их по временной переменной («Time») в соотношении:

- Train – 0.6
- Validation – 0.16
- Test – 0.23

2) Во втором наборе (BAF) из 1 000 000 транзакций 11 029 мошеннических. Всего 32 признака, присутствуют как числовые, так и категориальные переменные. Пропуски в данных отсутствуют. По результатам анализа распределений числовых переменных выполнено логарифмирование 8 переменных с тяжелыми хвостами распределений.

Для предобработки категориальных переменных использован One Hot Encoding. Далее выполнена конкатенация предобработанных категориальных и числовых переменных.

Для проведения моделирования на данном наборе данных будем разделять их по временной переменной («month») в соотношении:

- Train – 0.7
- Validation – 0.14
- Test – 0.14

После разделения выборки проведена нормализация данных.

4. Применение ML-методов и улучшение прогноза

4.1. Выбор метрик

В качестве метрик были выбраны стандартные метрики, включающие сравнение ожидаемой метки класса с меткой прогнозируемого класса, такие как:

- *Recall* (fraud/not fraud);

- *AUC*;

- *F1*

А также метрики для наборов данных с дисбалансом классов:

- *MCC* (1) – коэффициент корреляции Мэтьюса;

- *G-mean* (2) – среднее геометрическое Sensitivity (3) и Specificity(4)

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (1)$$

$$G - mean = \sqrt{Sensitivity * Specificity} \quad (2)$$

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

$$Specificity = \frac{TN}{(FP + TN)} \quad (4)$$

4.2. Построение базовых моделей на наборе данных CCF

CatBoost

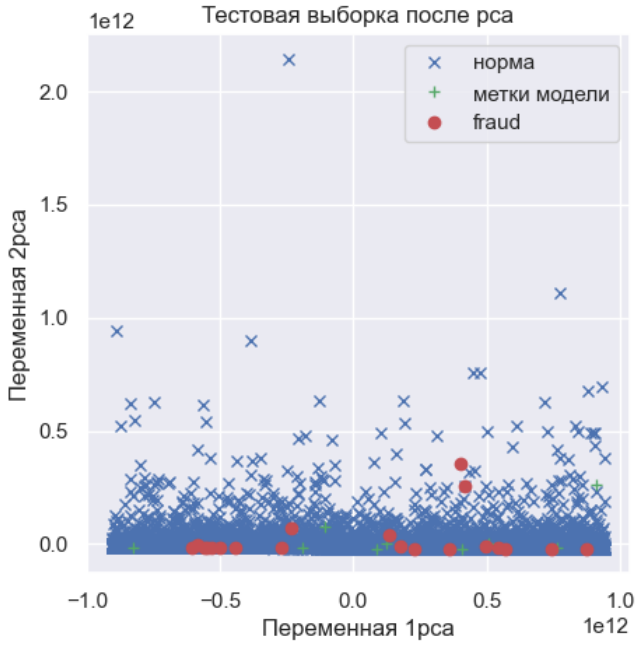

Рис. 4.2.1: Прогноз модели «CatBoost» для набора CCF (PCA 2)

LGBM

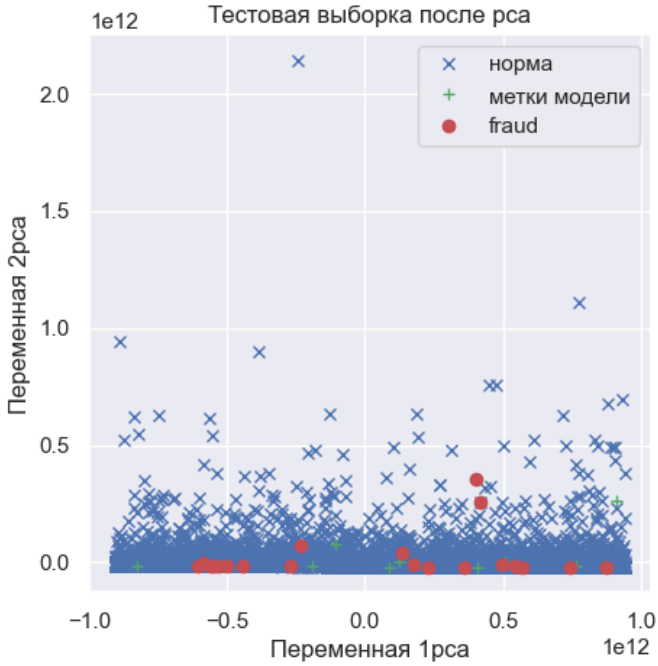

Рис. 4.2.2: Прогноз модели «LGBM» для набора CCF (PCA 2)

Результаты моделирования представлены в таблице 4.2.1.

	Recall not fr	**Recall fr**	**AUC**	**F1**	**MCC**	**G-mean**
CatBoost	0,99890	0,75	0,87445	0,56521	0,58257	0,86555
LGBM	**0,99979**	0,71153	0,85566	**0,75510**	**0,75624**	0,84343
HistGradientBoosting	0,99694	0,76923	0,88308	0,35874	0,42309	0,87571
DecisionTree	0,99757	0,71153	0,85455	0,38341	0,43108	0,84250
RandomForest	0,99925	0,76923	0,88424	0,64516	0,65323	0,87673
LogReg	0,99657	0,82692	**0,91174**	0,35537	0,43150	**0,90779**
MLP	0.00006	**1**	0,50003	0,00241	0,00029	0,00835
StackingClassifier	0,99194	0,82692	0,90943	0,19501	0,30057	0,90568
Iforest	0,87829	**0,90384**	0,89106	0,01765	0,08287	0,89097

Таблица 4.2.1: Значения метрик для различных моделей на наборе CCF

4.3. Построение базовых моделей на наборе данных BAF

CatBoost

Рис. 4.3.1: Прогноз модели «CatBoost» для набора BAF (PCA 2)

LGBM

Рис. 4.3.2: Прогноз модели «LGBM» для набора BAF (PCA 2)

Hist Gradient Boosting

Рис. 4.3.3: Прогноз модели «Hist Gradient Boosting» для набора BAF (PCA 2)

Logistic Regression

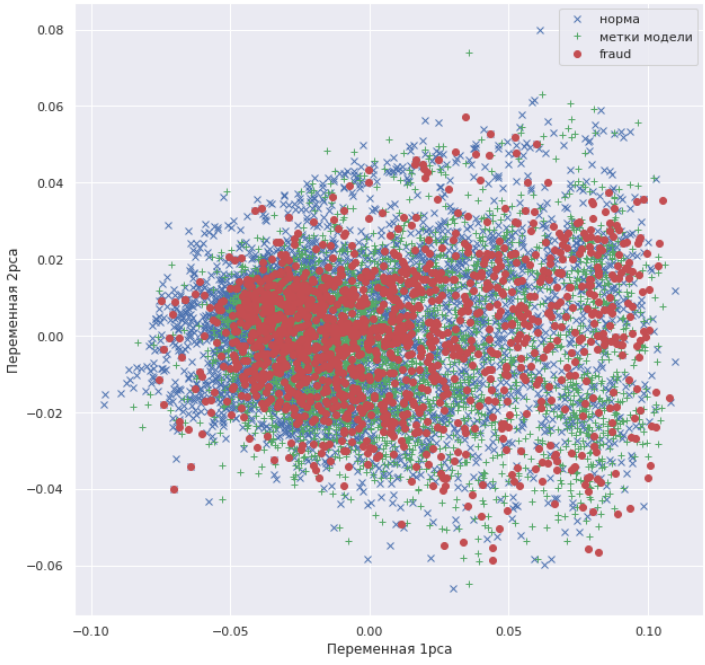

Рис. 4.3.4: Прогноз модели «Logistic Regression» для набора BAF (PCA 2)

Decision Tree

Рис. 4.3.5: Прогноз модели «Decision Tree» для набора BAF (PCA 2)

Random Forest

Рис. 4.3.6: Прогноз модели «Random Forest» для набора BAF (PCA 2)

Multi-layer Perceptron classifier (MLP)

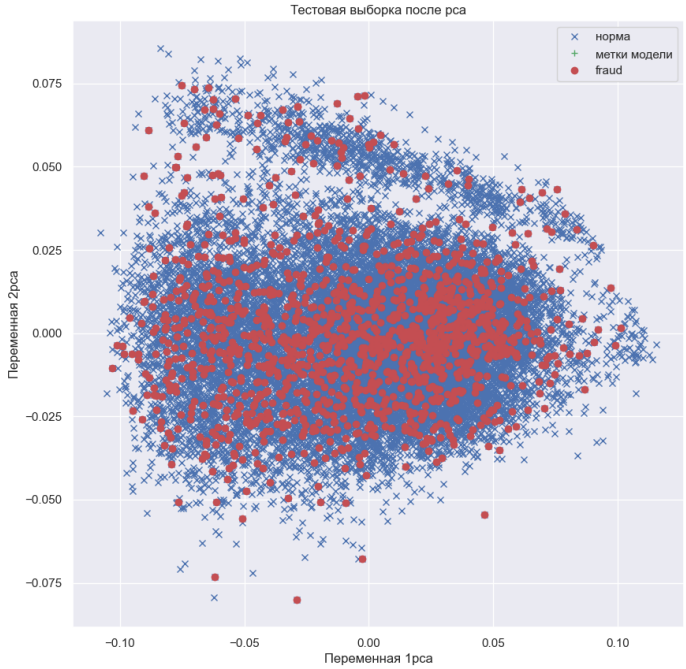

Рис. 4.3.7: Прогноз модели «MLP» для набора BAF (PCA 2)

Stacking Classifier (Decision Tree, Random Forest, Hist Gradient Boosting)

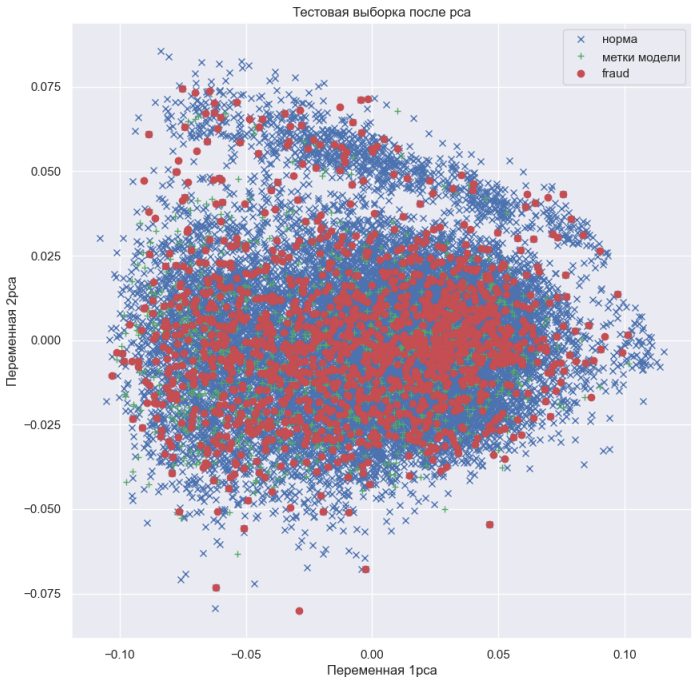

Рис. 4.3.8: Прогноз модели «Stacking Classifier» для набора BAF (PCA 2)

Одноклассовый SVM

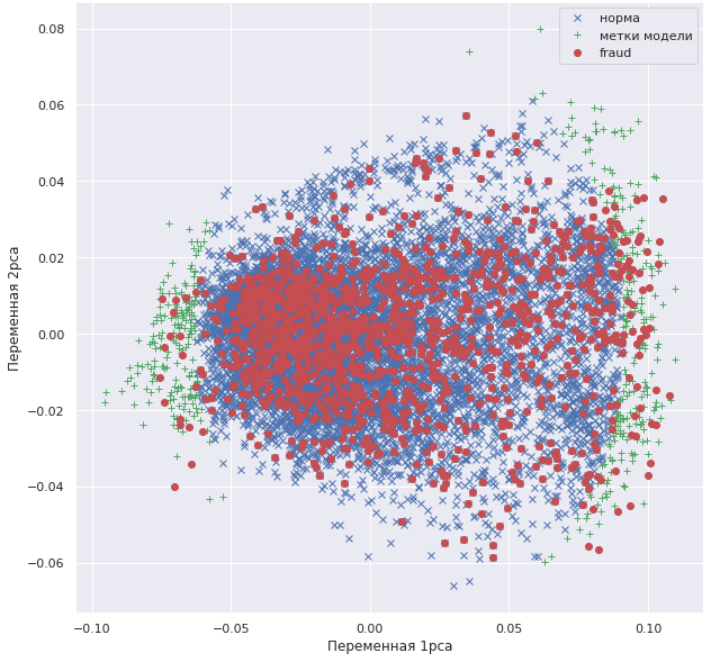

Рис. 4.3.9: Прогноз модели «Одноклассовый SVM» для набора BAF (PCA 2)

Isolation Forest

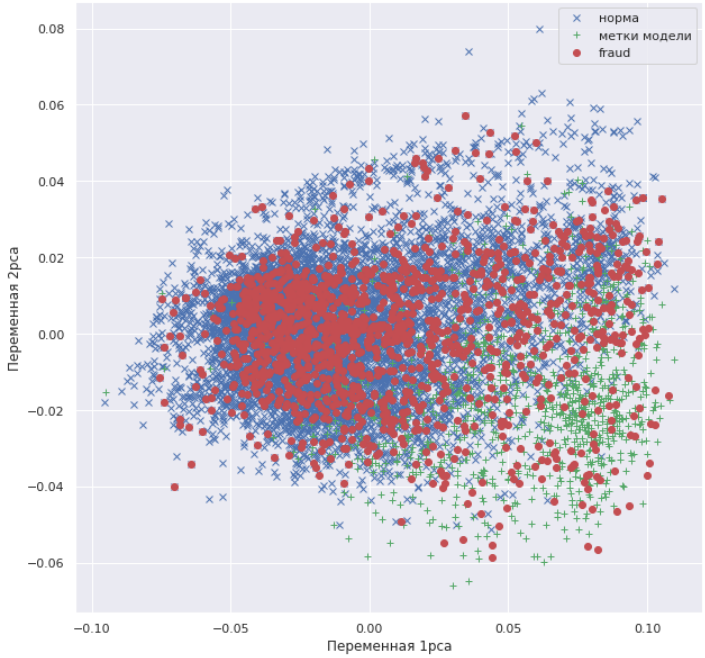

Рис. 4.3.10: Прогноз модели «Isolation Forest» для набора BAF (PCA 2)

Результаты моделирования представлены в таблице 4.1.

	Recall not fr	**Recall fr**	**AUC**	**F1**	**MCC**	**G-mean**
CatBoost	0,834683105	0,759207783	**0,796945**	0,113568	0,184446	**0,796051**
LGBM	**0,978781297**	0,255733148	0,617257	**0,186258**	0,178429	0,500307
HistGradientBoosting	0,853868493	0,72237665	0,788123	0,120539	0,187713	0,785375
DecisionTree	0,848218747	0,584086171	0,716152	0,095406	0,139495	0,70387
RandomForest	0,761043471	0,784225156	0,772634	0,084475	0,148827	0,772547
LogReg	0,274027497	**0,876997915**	0,575513	0,033179	0,039933	0,490226
MLP	**1**	0	0,5	0		0
StackingClassifier	0,946015742	0,461779013	0,703897	0,175805	**0,202478**	0,660946
Iforest	0,295127466	0,742876998	0,519002	0,028991	0,009808	0,468234
Одноклассовый svm	0,937120609	0,079221682	0,508171	0,028832	0,007907	0,272471

Таблица 4.3.1: Значения метрик для различных моделей на наборе BAF

4.4 Выводы

По результатам моделирования выбраны базовые модели для каждого набора данных.

В качестве базовой модели по метрике MCC для датасета CCF выбрана «LGBM» с результатом 0.756.

В качестве базовой модели по метрике MCC для датасета BAF выбран «Stacking Classifier» с результатом 0.202.

Также следует отметить, что хорошие результаты в наборе CCF показали модели CatBoost и LogReg. В наборе BAF хорошие результаты по метрике MCC показали модели HistGradientBoosting и CatBoost.

5. Практическое внедрение: DL-методы и доработка задачи

6. Заключение

Список литературы

[1] Abdelaty M. “Robust Anomaly Detection in Critical Infrastructure” B: The thesis 10.13140/RG.2.2.17844.94081 (2022). URL: https://www.researchgate.net/publication/364165605_Robust_Anomaly_Detection_in_Critical_Infrastructure (дата обр. 10.01.2023).

[2] Chater, M., Borgi, A., Slama M. “Fuzzy Isolation Forest for Anomaly Detection” B: Procedia Computer Science. 207. 916-925. 10.1016/j.procs.2022.09.147 (2022). URL: https://www.researchgate.net/publication/364464048_Fuzzy_Isolation_Forest_for_Anomaly_Detection (дата обр. 14.01.2023).

[3] Chen J, Zhang J, Qian R, Yuan J, Ren Y. “An Anomaly Detection Method for Wireless Sensor Networks Based on the Improved Isolation Forest.” B: Applied Sciences. 2023; 13(2):702. URL: https://doi.org/10.3390/app13020702 (дата обр. 10.03.2023).

[4] Denning, D.E. “An Intrusion Detection Model” B: IEEE Transactions on Software Engineering, Vol. SE-13, 222-232 (1987).

[5] Hongzuo Xu, Pang G., Wang Y. “Deep Isolation Forest for Anomaly Detection” B: 10.48550/arXiv.2206.06602 (2022). URL: https://www.researchgate.net/publication/361301387_Deep_Isolation_Forest_for_Anomaly_Detection(дата обр. 12.02.2023).

[6] Jerez C. I., Zhang J., Silva M. R. “On Equivalence of Anomaly Detection Algorithms” B: ACM Transactions on Knowledge Discovery from Data. 17. 10.1145/3536428 (2022). URL: https://www.researchgate.net/publication/360695739_On_Equivalence_of_Anomaly_Detection_Algorithms (дата обр. 02.03.2023).

[7] Kanishima Y., Sudo T., Yanagihashi H. “Autoencoder with Adaptive Loss Function for Supervised Anomaly Detection” B: Conference: Proc. of 26th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES2022) (2022). URL: https://www.researchgate.net/publication/363857063_Autoencoder_with_Adaptive_Loss_Function_for_Supervised_Anomaly_Detection (дата обр. 04.03.2023).

[8] Lok, L.K., Hameed V. A. Rana, M. E. “Hybrid Machine Learning Approach for Anomaly Detection” Indonesian Journal of Electrical Engineering and Computer Science. 27. 1016. 10.11591/ijeecs.v27.i2.pp1016-1024 (2022). URL: https://www.researchgate.net/publication/362400043_Hybrid_machine_learning_approach_for_anomaly_detection (дата обр. 20.03.2023).

[9] Sabeti E., Oh S., Song PXK, Hero AO. “A Pattern Dictionary Method for Anomaly Detection” B: Entropy 2022; 24(8):1095. URL: https://doi.org/10.3390/e24081095 (дата обр. 20.02.2023).

[10] Samariya D., Thakkar A. “A Comprehensive Survey of Anomaly Detection Algorithms” B: Annals of Data Science (2021). URL: https://link.springer.com/article/10.1007/s40745-021-00362-9 (дата обр. 25.02.2023).

[11] H. S. Teng, K. Chen and S. C. Lu, "Adaptive real-time anomaly detection using inductively generated sequential patterns," B: Proceedings. 1990 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, CA, USA, 1990, pp. 278-284, doi: 10.1109/RISP.1990.63857.

[12] Thudumu S., Branch Ph., Jin J. “A comprehensive survey of anomaly detection techniques for high dimensional big data” B: Journal of Big Data. 7. 10.1186/s40537-020-00320-x. (2020). URL: https://www.researchgate.net/publication/342638066_A_comprehensive_survey_of_anomaly_detection_techniques_for_high_dimensional_big_data (дата обр. 15.01.2023).