ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Направление подготовки 01.04.02 Прикладная математика и информатика

Образовательная программа

«Машинное обучение и высоконагруженные системы»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Детектирование аномалий в данных

Подготовили:
Кулакин Сергей
Журович Янина

Руководитель:
Максимовская Анастасия

Москва, 2023

Оглавление

Глава 1. Введение

Начать стоит с того, что под аномалиями понимаются некоторые необычности части данных, выбросы, шумы, отклонения или исключения. Если немного углубиться, то, строго говоря, в анализе данных существуют два направления, которые занимаются поиском аномалий: *детектирование выбросов* (outlier detection) и *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 как раз и 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 «боретесь 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 нижеследующих датасета:

- 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 (492 мошенничества из 284 807 транзакций за период двух дней). К его особенностям можно отнести ощутимую несбалансированность: так, на положительный класс (мошенничество) приходится 0.172% всех транзакций. Отсюда следует рекомендация по измерению точности с помощью показателя AUPRC – площади под кривой точности отзыва. 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.
- https://www.kaggle.com/datasets/ealaxi/banksim1 — более редко используемый датасет и выглядит интереснее с той точки зрения, что в данном случае применялись синтетические данные. Каким образом это было сделано: существует BankSim — агентский симулятор банковских платежей, основанный на выборке агрегированных данных о транзакциях, предоставленных банком в Испании. Основная его цель — генерация синтетических данных, которые можно использовать для исследований по обнаружению мошенничества. В данном случае использовались статистический анализ и анализ социальных сетей (SNA), а именно отношений между продавцами и покупателями, чтобы разработать и откалибровать модель. Цель создателей состоит в том, чтобы BankSim можно было использовать для моделирования соответствующих сценариев, сочетающих обычные платежи и внедрение известных мошеннических сигнатур. При этом, наборы данных, созданные BankSim, не содержат личной информации или раскрытия юридических и частных транзакций клиентов. Таким образом, научное

сообщество и другие организации могут использовать его для разработки и обоснования методов обнаружения мошенничества. Синтетические данные имеют дополнительное преимущество, заключающееся в том, что их легче получить, быстрее и с меньшими затратами для экспериментов (даже для тех, у кого есть доступ к собственным данным). Кроме того, утверждается, что BankSim генерирует данные, которые достаточно приближаются к соответствующим аспектам реальных данных.

В данном случае, BankSim запускался на 180 шагов (примерно шесть месяцев) несколько раз, а также калибровались параметры, чтобы получить распределение, достаточно близкое, чтобы быть надежным для тестирования. Затем были собраны несколько лог-файлов и выбран наиболее точный. Помимо этого, были внедрены «воры», которые стремятся украсть в среднем три карты за шаг и совершить около двух мошеннических транзакций в день. При этом, всего в итоге было собрано 594643 единиц данных, где 587443 — это обычные платежи и 7200 мошеннических транзакций. Поскольку это рандомизированное моделирование, то значения, конечно, не идентичны исходным данным.

– 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 (df.describe(include='all'));
- 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

– LogReg + CV

– CatBoost

– Semi-supervised

– GAN

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.

Глава 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

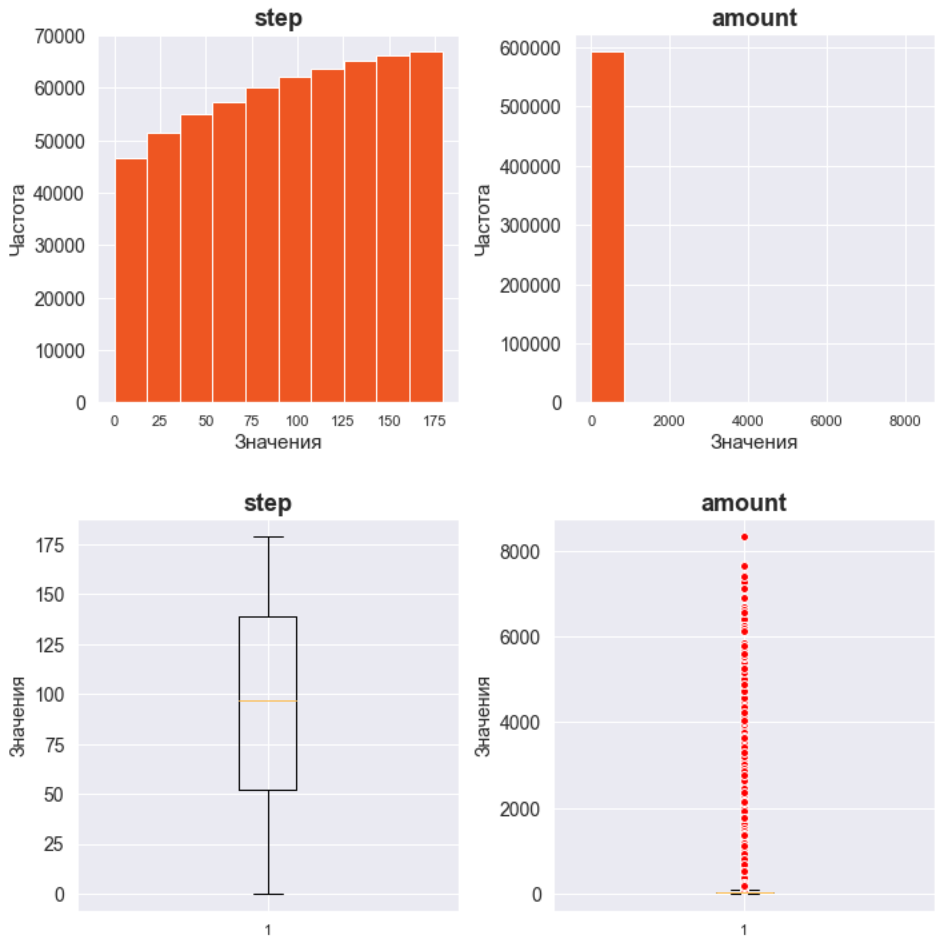
метод превосходит стандартный алгоритм Isolation Forest с двумя выбранными в данном случае наборами данных.

Глава 3. Разведочный анализ данных и моделирование

В данном случае разведочный анализ данных включает в себя процесс отбора вышеупомянутых датасетов для исследования, так как изначально было выбрано 6 наборов данных, а осталось только 2 наиболее подходящих.

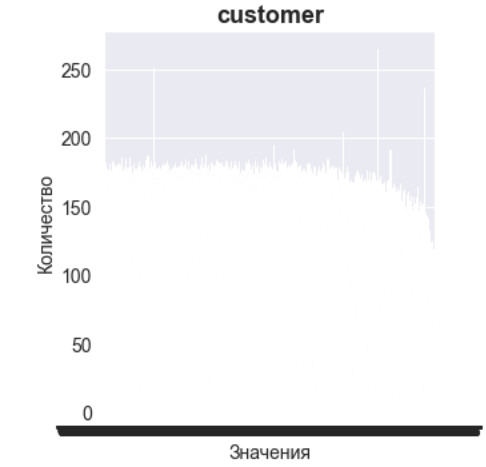
1. Synthetic data from a financial payment system
   https://www.kaggle.com/datasets/ealaxi/banksim1
   – анализ основных характеристик данных в том числе показал, что всего **594643** наблюдений, и только **7200 мошеннических** транзакции, то есть присутствует некоторый дисбаланс:

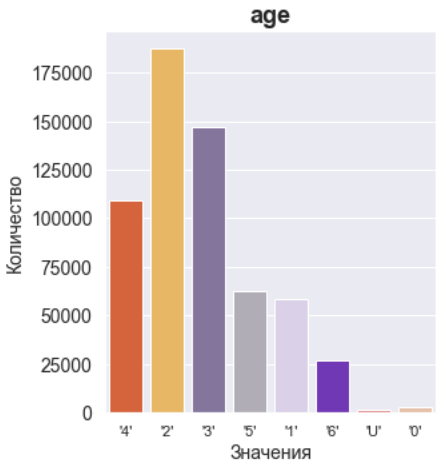
   – были опробованы описательные статистики, а также обнаружено отсутствие пропусков в данных обеих таблиц, при это выявлен факт того, что некоторые столбцы дублируются и не очень понятно, зачем это сделано;
   – помимо этого был реализован некий анализ числовых переменных, в ходе которого выяснилось, что числовая переменная суммы (amount) имеет значительное количество выбросов:

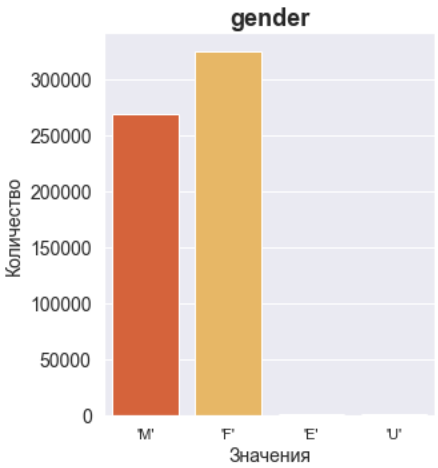


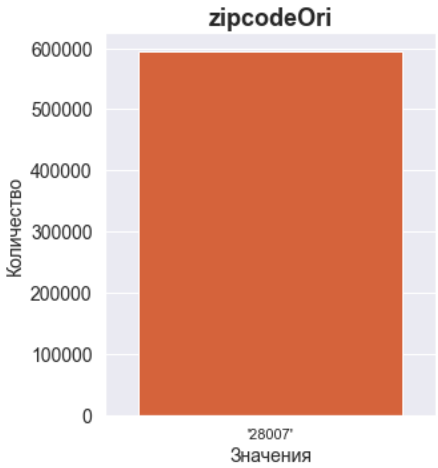
– кроме того, были рассмотрены категориальные переменные:

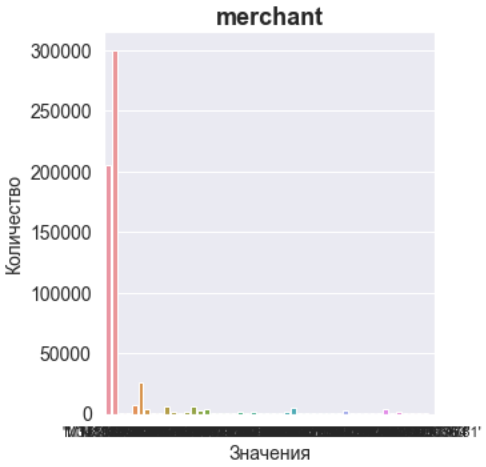
- zipcodeOri, zipMerchant не несут полезной информации, в каждой по одному уникальному значению;
- Переменная age (возраст) имеет только 8 уникальных значений (скорее это возрастные группы);
- Переменная gender (пол) имеет 4 значения;
- Переменные merchant (торговец) и category (категория) имеют большое количество редких классов.

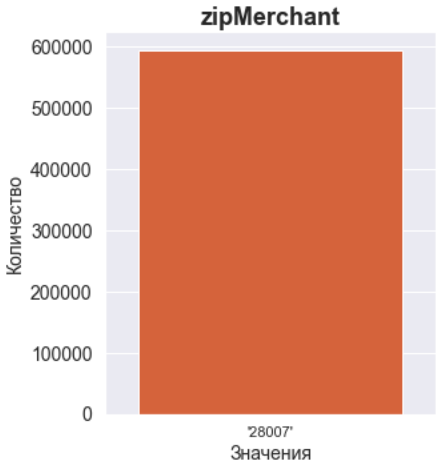

customer
Количество
Значения

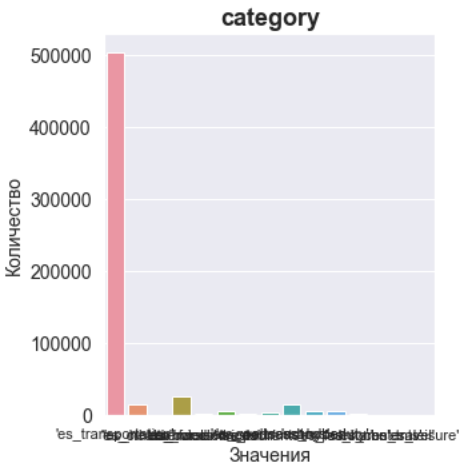

age
Количество
'4'
'2'
'3'
'5'
'1'
'6'
'U'
'0'
Значения

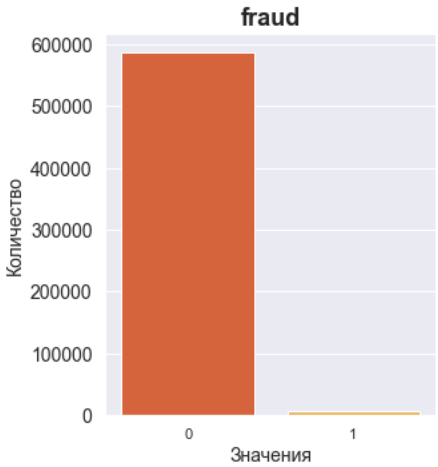

gender
Количество
'M'
'F'
'E'
'U'
Значения


zipcodeOri
Количество
'28007'
Значения


merchant
Количество
Значения

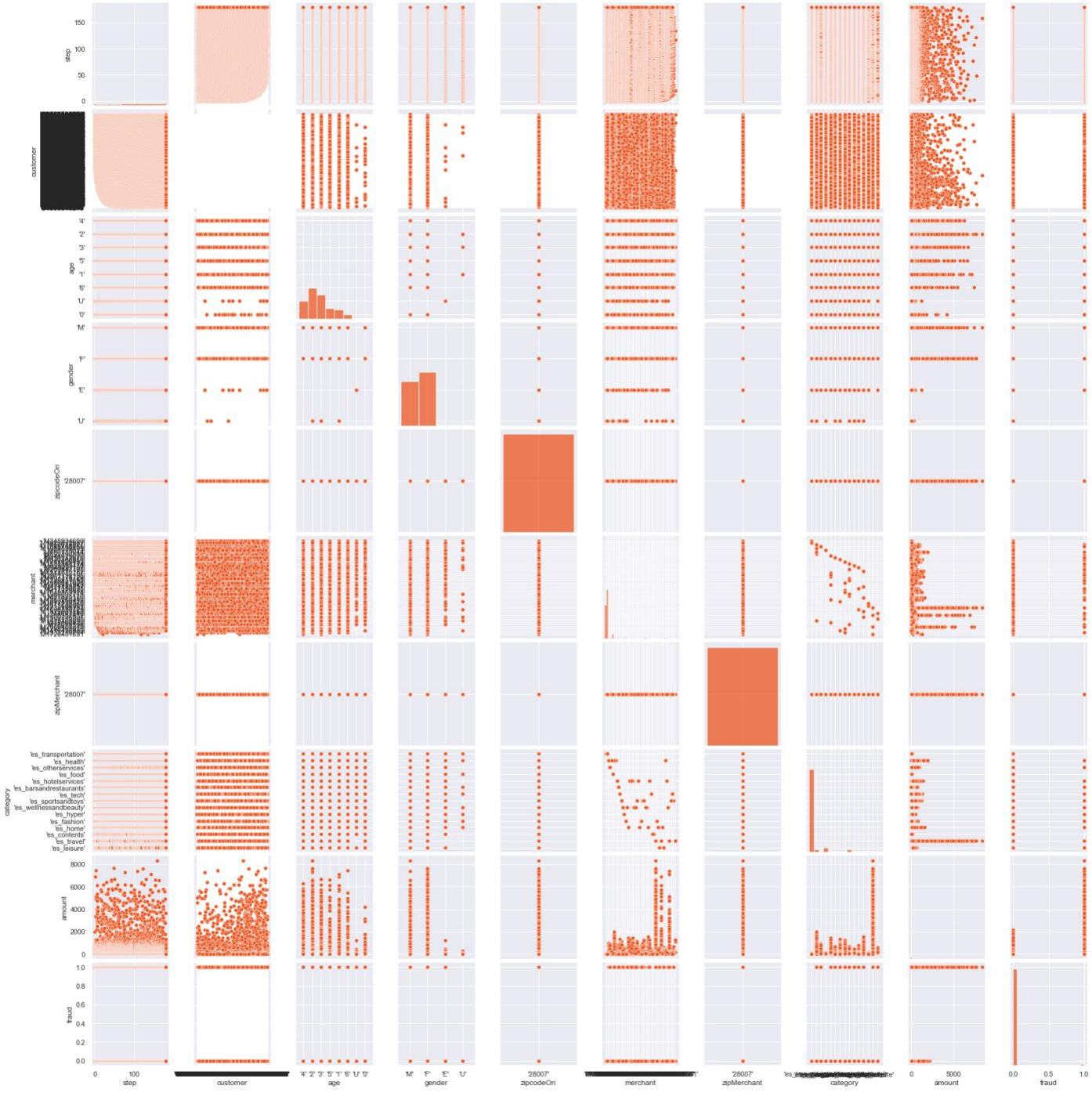

zipMerchant
Количество
'28007'
Значения


category
Количество
Значения

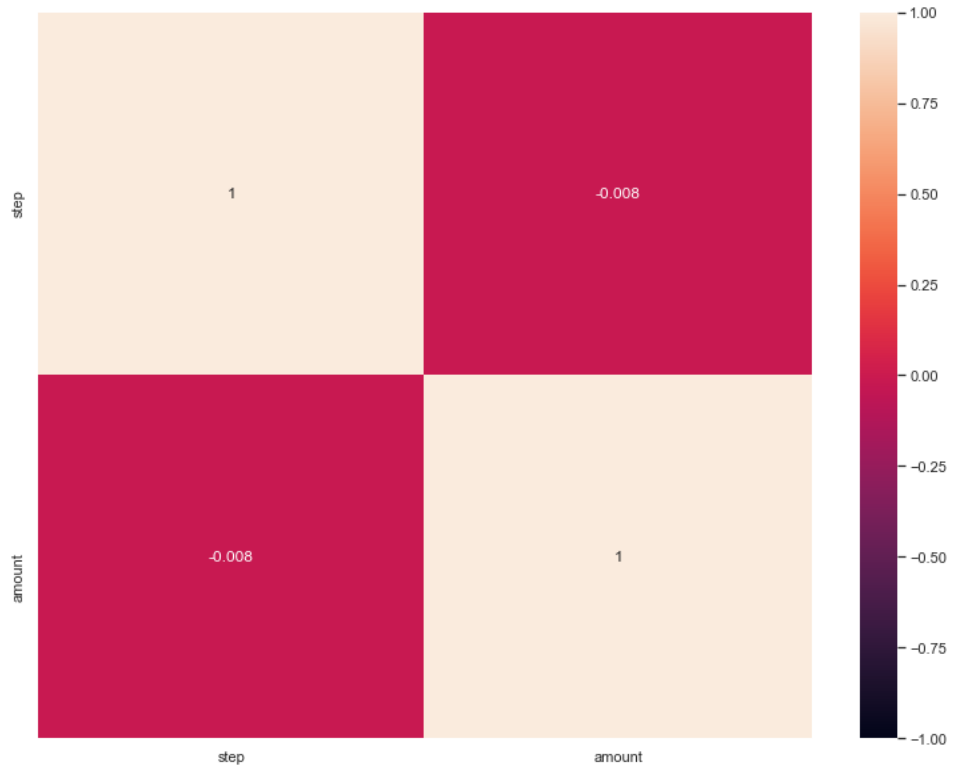

fraud
Количество
0
1
Значения

– парные распределения с таргетом:

- Переменные amount, category, merchant визуально взаимосвязаны с переменной fraud



– матрица корреляций числовых признаков демонстрирует незначительные показатели и выглядит следующим образом:



плюсы	минусы
нет проблемы приватности данных	не самое большое число признаков для моделирования
	искусственное происхождение данных

Подытог: такой датасет может быть использован для данного исследования.

2. Bank Transaction Data https://www.kaggle.com/datasets/apoorvwatsky/bank-transaction-data

   – данные результатов разведочного анализа показали, что всего в датасете представлена 116201 транзакция;

   – данные не являются размеченными, что является значительным недостатком;

   – данные выглядят как открытые ("консолидированные и извлеченные выписки по банковским счетам различных банковских счетов");

   Подытог: не приветствуется использование подобного датасета в исследовании.

3. Fraudulent Transactions Data https://www.kaggle.com/datasets/chitwanmanchanda/fraudulent-transactions-data

– в ходе EDA был выявлен тот факт, что всего в этом наборе данных присутствуют 5592 транзакции, что крайне мало;

– данные снова не являются размеченными, и это определённо минус;

– при этом данные открытые, и взяты они из ежемесячной статистики RBI: https://www.rbi.org.in/scripts/ATMView.aspx

Подытог: по сумме всех факторов было решено не использовать такой датасет.

4. Fraudulent Transactions Data
   https://www.kaggle.com/datasets/chitwanmanchanda/fraudulent-transactions-data

   – датасет состоит из 6362620 транзакций, но только 8213 из них являются мошенническими;

   – в данных не было обнаружено пропусков;

   – что же касается числовых переменных:

   - только переменная step не имеет большого числа выбросов;
   - все остальные числовые переменные ('amount', 'oldbalanceOrg', 'newbalanceOrig', 'oldbalanceDest', 'newbalanceDest') с большим количеством выбросов;

   – насчёт категориальных переменных можно утверждать следующее:

   - переменная nameOrig содержит практически уникальные значения;
   - nameDest и nameOrig имеют очень много уникальных значений (скорее всего это id);
   - type имеет 5 значений;

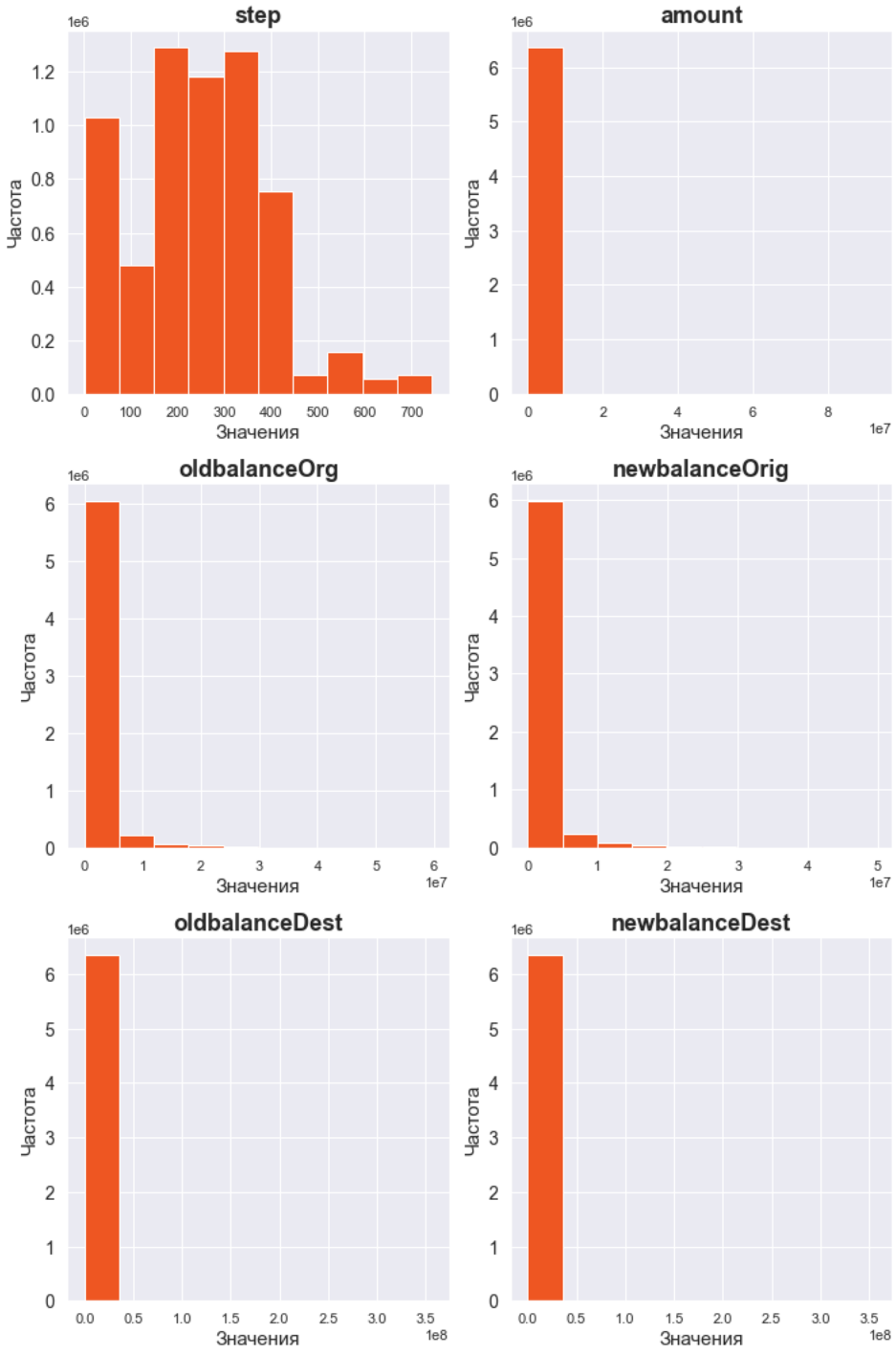
   – парные распределения признаков показывают взаимосвязь отдельных переменных с целевой (таргетом);

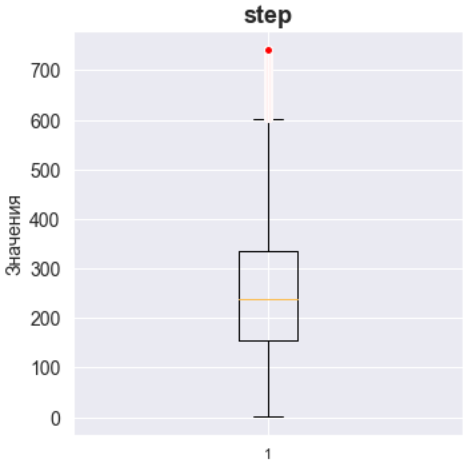
   – корреляции числовых переменных прослеживаются в основном с переменной amount;

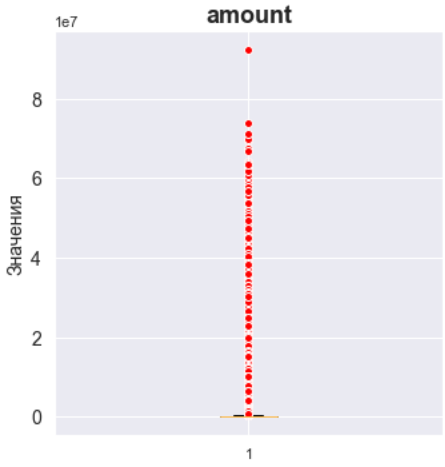
плюсы	минусы
достаточное количество признаков, которые коррелируют с таргетом	сомнительный источник данных

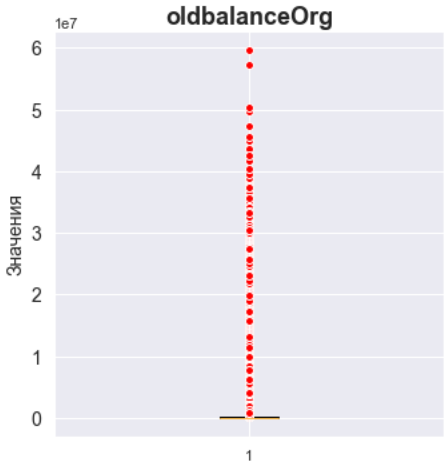
	отсутствие корректного и достаточно полного описания набора данных

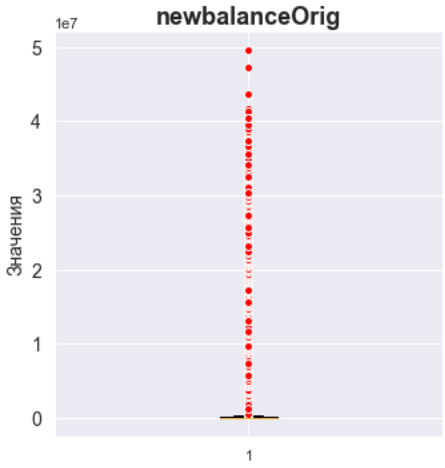
** далее следует небольшой раздел-приложение с подтверждающими выводы графиками и диаграммами:*

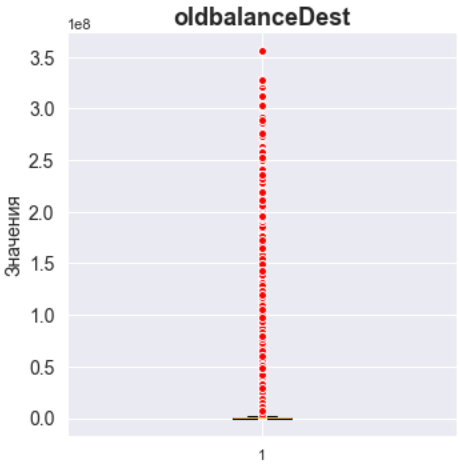


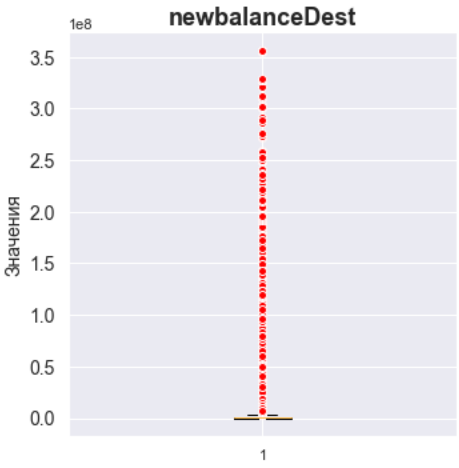

step
Значения
700
600
500
400
300
200
100
0
1

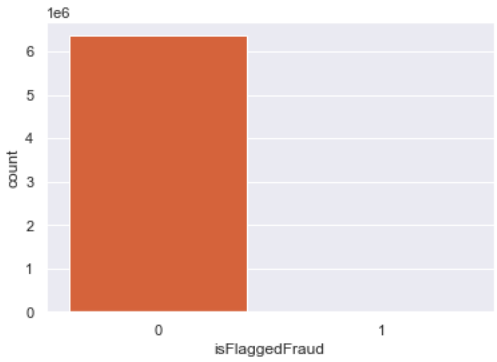

amount
1e7
Значения
8
6
4
2
0
1

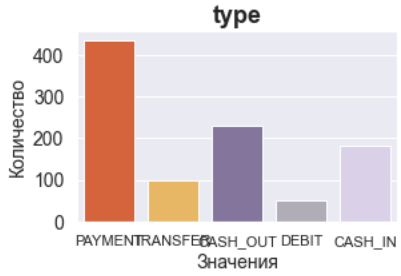

oldbalanceOrg
1e7
Значения
6
5
4
3
2
1
0
1


newbalanceOrig
1e7
Значения
5
4
3
2
1
0
1

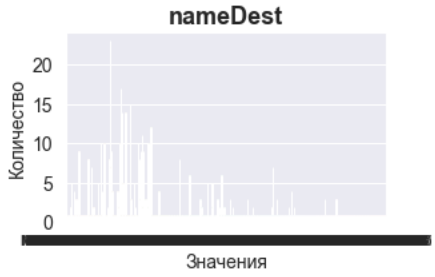

oldbalanceDest
1e8
Значения
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1


newbalanceDest
1e8
Значения
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

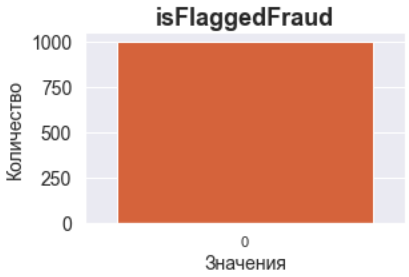

1e6
count
6
5
4
3
2
1
0
0
1
isFlaggedFraud


type
Количество
400
300
200
100
0
PAYMENT
TRANSFER
CASH_OUT
DEBIT
CASH_IN
Значения

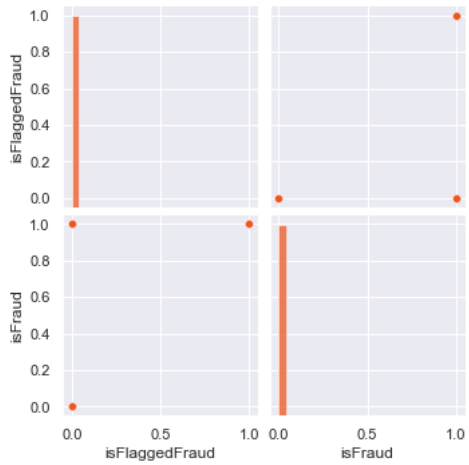

nameOrig
Количество
1.00
0.75
0.50
0.25
0.00
Значения

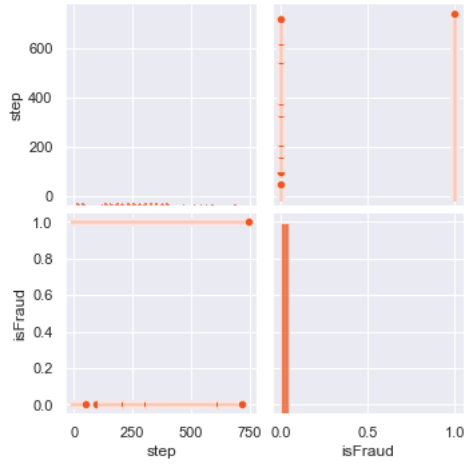

nameDest
Количество
20
15
10
5
0
Значения

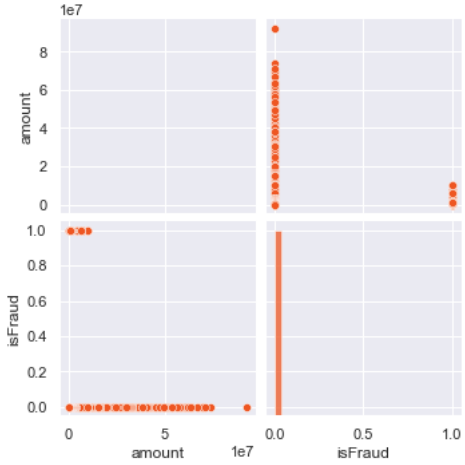

isFraud
Количество
1000
750
500
250
0
0
1
Значения

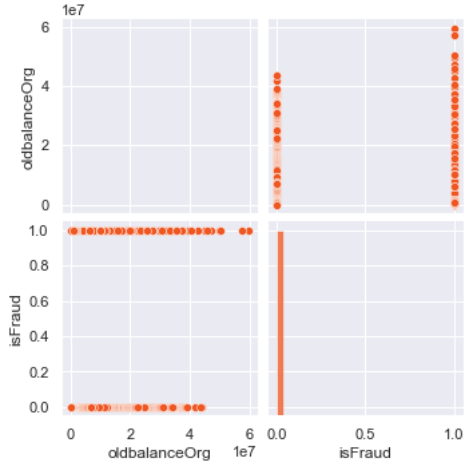

isFlaggedFraud
Количество
1000
750
500
250
0
0
Значения

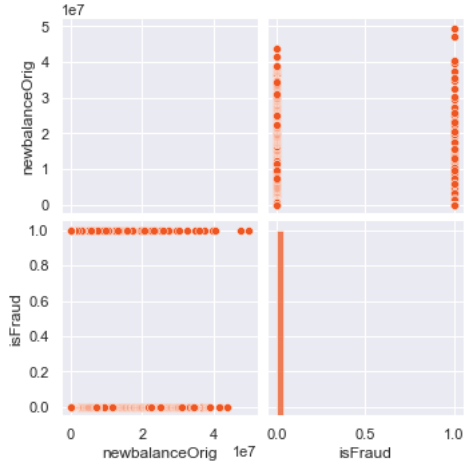

type
PAYMENT
TRANSFER
CASH_OUT
DEBIT
CASH_IN
isFraud
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0

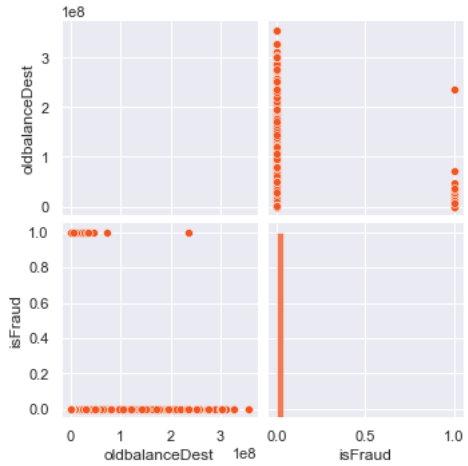

isFlaggedFraud
1.0
0.8
0.6
0.4
0.2
0.0
isFraud
0.0
0.5
1.0

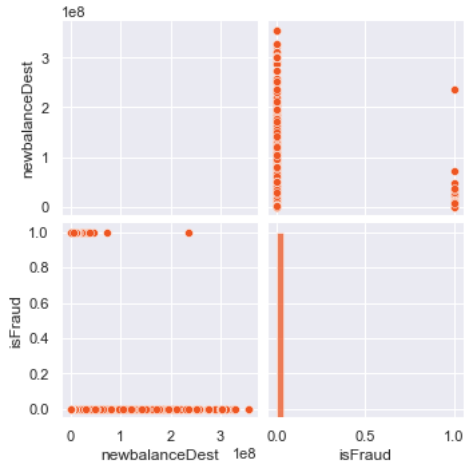

step
isFraud
0
250
500
750
0.0
0.5
1.0


1e7
amount
isFraud
0
5
1e7
0.0
0.5
1.0


1e7
oldbalanceOrg
isFraud
0
2
4
6
1e7
0.0
0.5
1.0


1e7
newbalanceOrig
isFraud
0
2
4
1e7
0.0
0.5
1.0


1e8
oldbalanceDest
isFraud
0
1
2
3
1e8
0.0
0.5
1.0


1e8
newbalanceDest
isFraud
0
1
2
3
1e8
0.0
0.5
1.0

Подытог: предложенный датасет может быть применён в рамках данной работы.

5. Fraudulent Transactions Prediction
   https://www.kaggle.com/datasets/vardhansiramdasu/fraudulent-transactions-prediction
   – по сути, является аналогией или даже копией предыдущего рассмотренного датасета;
   – исходя из предыдущего пункта все выводы дублируются;
   Подытог: его также можно рассмотреть на предмет применения в исследовании.

6. Credit Card Fraud Detection https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download
   – небольшая ремарка в виде выдержки из описания этого популярного датасета: Он содержит только числовые входные переменные, которые являются результатом преобразования PCA. К сожалению, из-за проблем с конфиденциальностью мы не можем предоставить исходные функции и дополнительную справочную информацию о данных. Функции V1, V2, ... V28 являются основными компонентами, полученными с помощью PCA, единственными функциями, которые не были преобразованы с помощью PCA, являются "Время" и "Количество". Функция

"Время" содержит секунды, прошедшие между каждой транзакцией и первой транзакцией в наборе данных. Функция "Сумма" – это сумма транзакции, эта функция может быть использована, например, для обучения, зависящего от затрат. Функция 'Class' является переменной ответа и принимает значение 1 в случае мошенничества и 0 в противном случае;

– датасет содержит в себе всего **284807** и только **492** мошеннических транзакции;

– все признаки в выборке являются числовыми;

– помимо этого, в выборке отсутствуют пропуски;

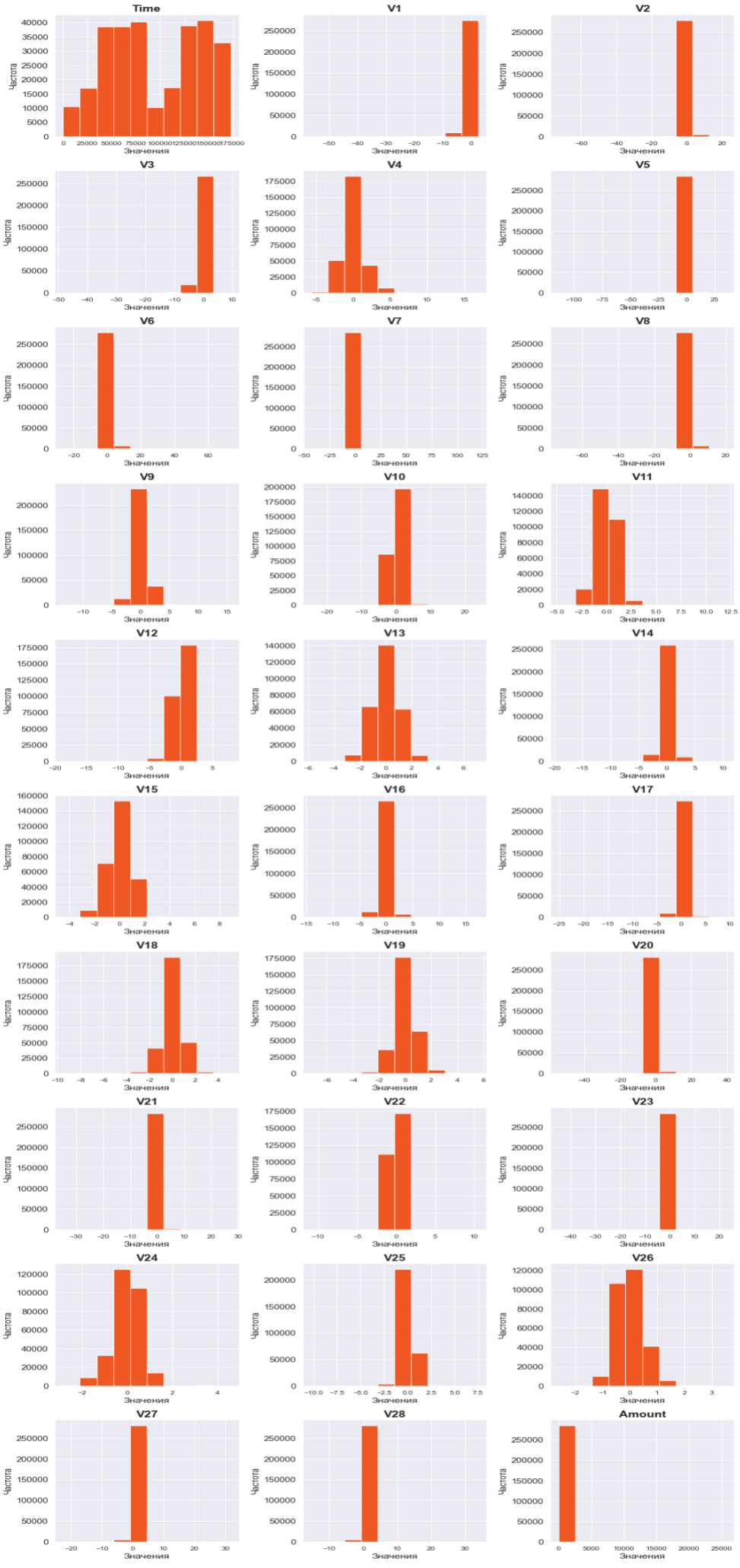
– при рассмотрении категориальных переменных стало понятно, что все признаки кроме времени имеют большое количество выбросов;

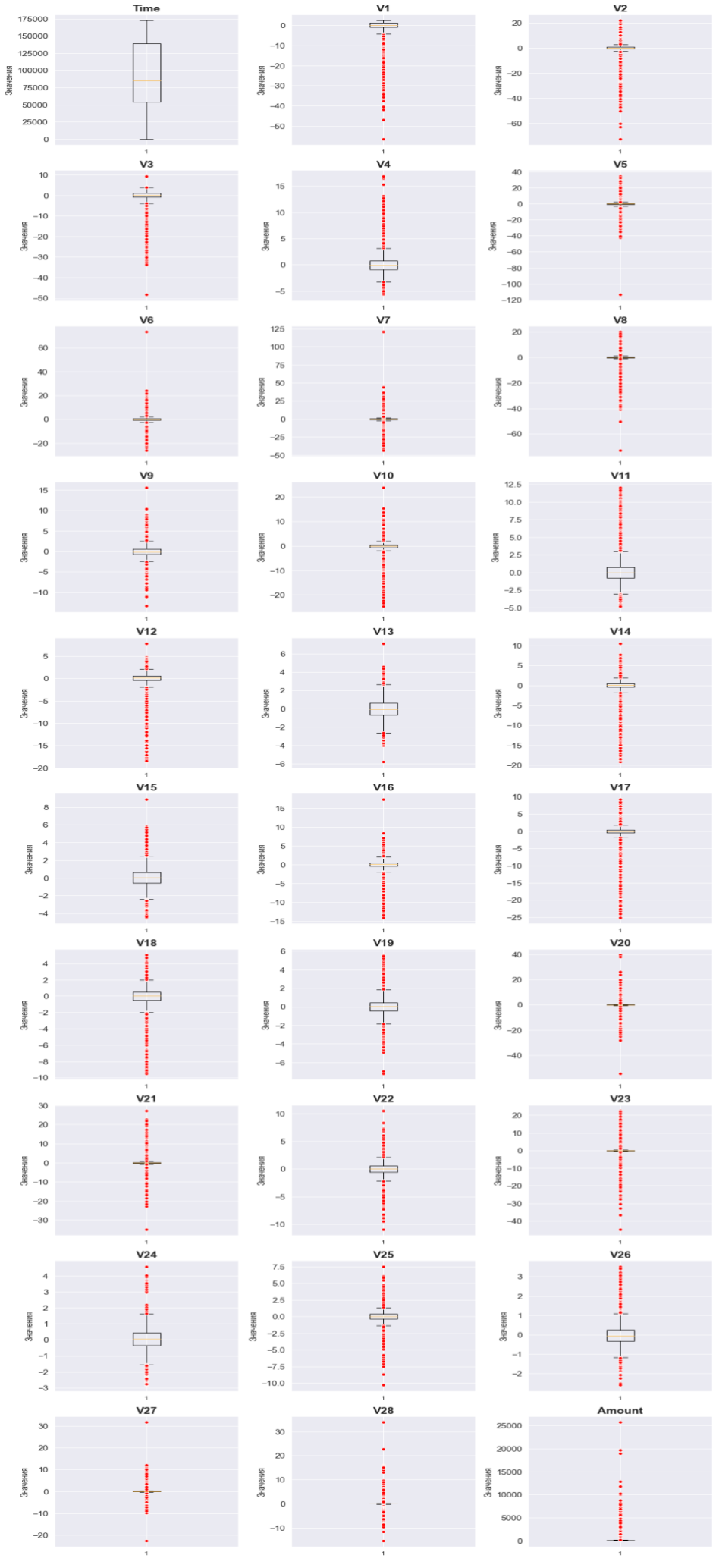
– парные распределения переменных с таргетом наглядно доказывают наличие корреляции между ними;

– корреляционная матрица числовых признаков подтверждает наличие корреляции с Class;

плюсы	минусы
достаточное количество признаков, коррелирующих с целевой переменной	высокая популярность и используемость данного датасета
	наличие большого количества выбросов, которые нуждаются в анализе и обработке
	для сохранения приватности данные содержат главные компоненты вместо исходных признаков транзакций, за исключением ‘Time’ и ‘Amount’

Подытог: по совокупности всех критериев решено использовать этот датасет.
** далее следует небольшой раздел-приложение с подтверждающими выводы графиками и диаграммами:*


Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Частота
Значения


Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Значения
1

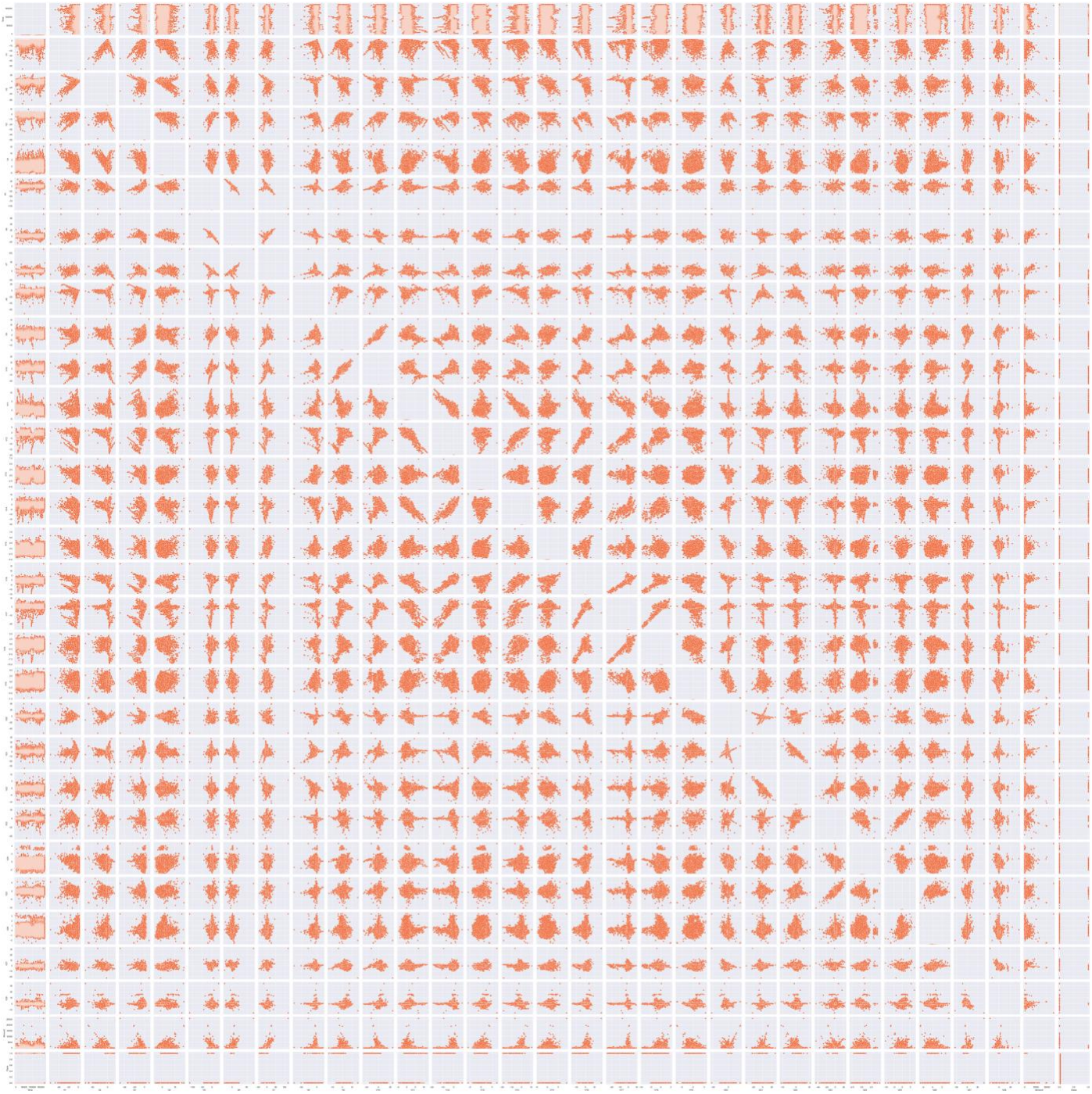
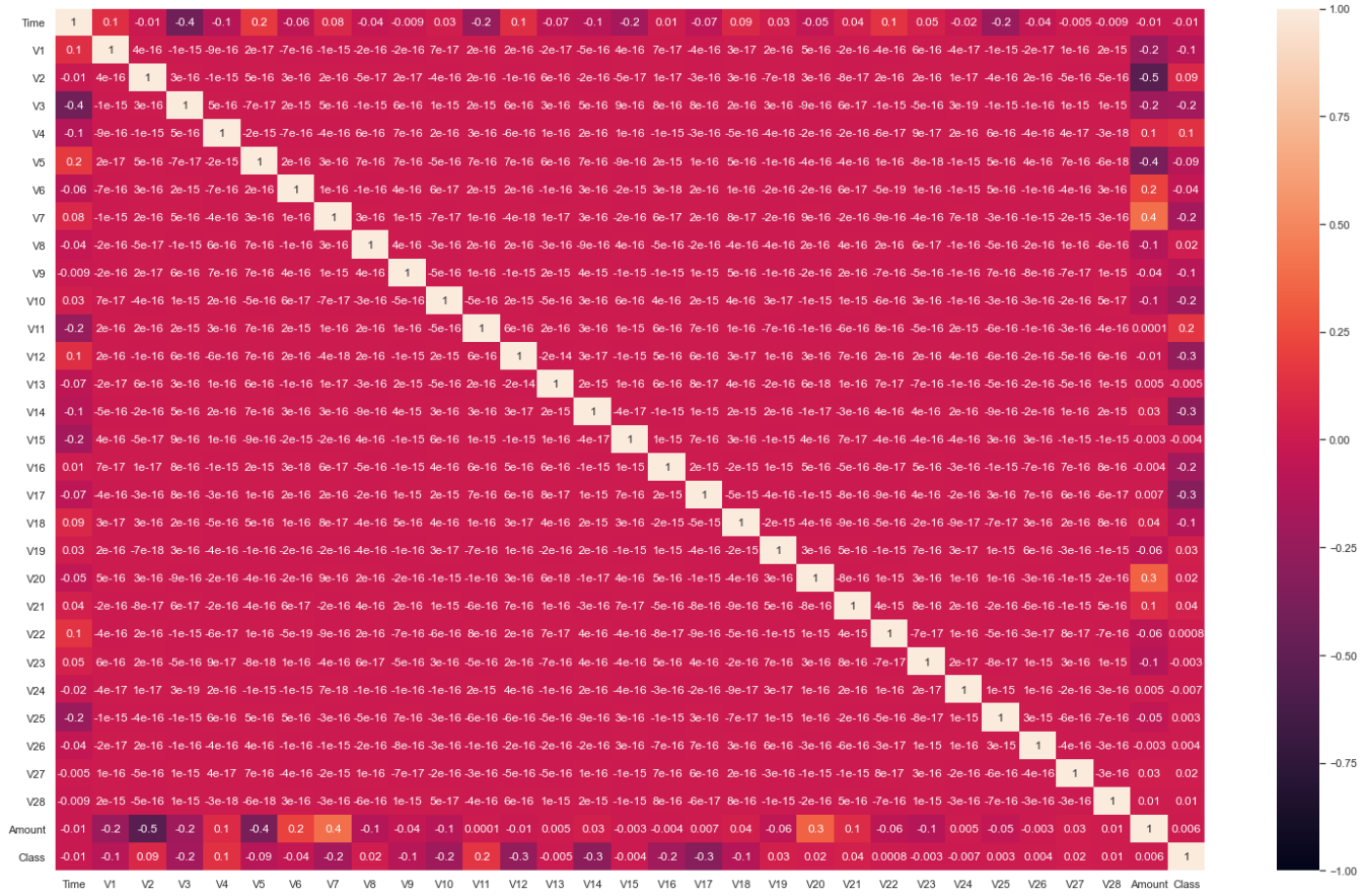
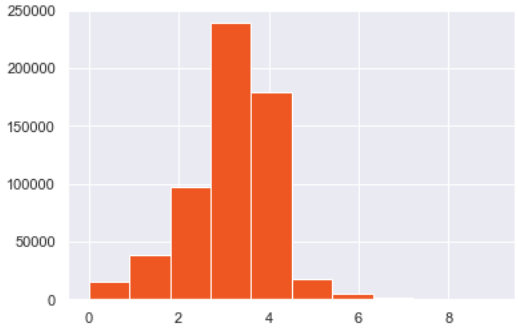
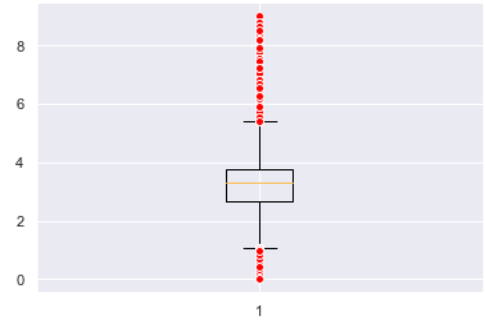



График соотношения классов
250000
200000
150000
100000
50000
0
Количество
fraud
not_fraud
Класс



На данный момент готово моделирование по первому из выбранных датасету. В дальнейшем на всех этапах планируется использовать оба набора данных. Итак, первым этапов была реализована подготовка числовых данных. Исходя из выводов EDA, была в основном рассмотрена переменная ‘Amount’, так как именно для неё наблюдались выбросы:





Было принято решение, что обрезать данные по квантилю не стоит, потому что на уровне 0.99 квантиля порядка 5К фрода. Затем была проведена подготовка категориальных переменных: сначала были удалены неинформативные колонки, а

потом исследованы переменные ‘Age’, ‘Gender’, ‘Merchant’, ‘Category’. После этого этапа было произведено избавление от дисбаланса в классах:

В ходе чего размер выборки увеличился в два раза. Далее были выбраны метрики для последующей оценки качества моделей, а именно: accuracy, F1-score, ROC-AUC. После этого был запущен и реализован процесс стандартизации и подготовки данных, в процессе которого было применено one-hot кодирование. Затем была обучена линейная модель логистической регрессии с применением кросс-валидации (значение точности на тестовой выборке приблизительно равно 0.991, f-меры – 0.991, а ROC-AUC ~ 0.9983). В конце ещё была применена модель CatBoost, все те же показатели качества на которой были равны соответственно ~ 0.9992.

Глава 4. Применение ML-методов и улучшение прогноза

Глава 5. Практическое внедрение: DL-методы и доработка задачи

Глава 6. Заключение

Список использованной литературы и других источников

1. Abdelaty M. (2022). Robust Anomaly Detection in Critical Infrastructure // The thesis: https://www.researchgate.net/publication/364165605_Robust_Anomaly_Detection_in_Critical_Infrastructure
2. *Chater, M., Borgi, A., Slama M. T. [and etc.]* (2022). Fuzzy Isolation Forest for Anomaly Detection // Procedia Computer Science, 207: 916-925. https://www.researchgate.net/publication/364464048_Fuzzy_Isolation_Forest_for_Anomaly_Detection
3. *Chen J., Zhang J., Qian R., [and etc.]* (2023). An Anomaly Detection Method for Wireless Sensor Networks Based on the Improved Isolation Forest // Appl. Sci. *2023*, 13*(2)*, *702*; https://www.mdpi.com/2076-3417/13/2/702
4. *Denning, D. E.* (1987). An Intrusion Detection Model // IEEE Transactions on Software Engineering, Vol. SE-13, 222-232.
5. Hongzuo Xu, Pang G., Wang Y. (2022). Deep Isolation Forest for Anomaly Detection // https://www.researchgate.net/publication/361301387_Deep_Isolation_Forest_for_Anomaly_Detection.
6. *Jerez C. I., Zhang J., Silva M. R.* (2022). On Equivalence of Anomaly Detection Algorithms // ACM Transactions on Knowledge Discovery from Data https://www.researchgate.net/publication/360695739_On_Equivalence_of_Anomaly_Detection_Algorithms
7. *Kanishima Y., Sudo T., Yanagihashi H.* (2022). Autoencoder with Adaptive Loss Function for Supervised Anomaly Detection // Conference: Proc. of 26th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES2022). https://www.researchgate.net/publication/363857063_Autoencoder_with_Adaptive_Loss_Function_for_Supervised_Anomaly_Detection
8. *Lok, L.K., Hameed V. A.* (2022). Hybrid Machine Learning Approach for Anomaly Detection // Indonesian Journal of Electrical Engineering and Computer Science 27(2): 1016. https://www.researchgate.net/publication/362400043_Hybrid_machine_learning_approach_for_anomaly_detection , datasets: “Index of /datasets.”

[Online]. packages.revolutionanalytics.com. http://packages.revolutionanalytics.com/datasets/ (accessed May 16, 2021).

9. *Sabeti E., Sehong Oh, Peter X. K. Song* (2022). A Pattern Dictionary Method for Anomaly Detection // Entropy, 24, 1095. *Sabeti E., Sehong Oh, Peter X. K. Song* (2022). A Pattern Dictionary Method for Anomaly Detection // Entropy, 24, 1095. https://www.mdpi.com/1099-4300/24/8/1095
10. *Samariya D., Thakkar A.* (2021). A Comprehensive Survey of Anomaly Detection Algorithms // Annals of Data Science. https://link.springer.com/article/10.1007/s40745-021-00362-9
11. *Teng H. S., Chen K., Lu S. C.* (1990). Adaptive real-time anomaly detection using inductively generated sequential patterns // Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. — ISBN 978-0-8186-2060-7. — doi:10.1109/RISP.1990.63857.
12. *Thudumu S., Branch Ph., Jin J.* (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data // Journal of Big Data 7(1). https://www.researchgate.net/publication/342638066_A_comprehensive_survey_of_anomaly_detection_techniques_for_high_dimensional_big_data