ВЫЯВЛЕНИЕ ПОДОЗРИТЕЛЬНЫХ УЗЛОВ СЕТИ БИТКОИН МЕТОДАМИ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Гарин Л.А.[1], Гисин В.Б.[2]

Цель исследования: *проанализировать существующие модели машинного обучения, позволяющие выявлять подозрительные адреса сети Биткоин, разработать современную эффективную скоринговую модель для выявления подозрительных адресов.*

Метод исследования: *сбор и анализ данных об адресах и транзакциях сети Биткоин, выявление признаков подозрительной активности адресов, анализ, построение и экспериментальная проверка эффективности моделей машинного обучения, направленных на выявление подозрительных адресов сети Биткоин на основе транзакций их владельцев.*

Полученные результаты: *проведен анализ распространенных наборов данных и моделей машинного обучения, связанных с выявлением подозрительных адресов сети Биткоин, собраны данные о транзакциях, связанных с представительным набором адресов. Построены модели машинного обучения для выявления подозрительных адресов на основе собранной информации. Проведена экспериментальная апробация моделей. Установлено, что наилучший результат дает модель, использующая градиентный бустинг. Эта модель демонстрирует более эффективную работу по сравнению с имеющимися аналогами.*

Ключевые слова: *распределенный реестр, блокчейн, транзакция, скоринг, машинное обучение, криптовалюта, подозрительная активность.*

DOI: 10.21681/2311-3456-2022-3-90-99

Введение

Сегодня криптовалюту в некоторых странах признают средством платежа, а организации принимают в качестве оплаты; признают имуществом, тем самым устанавливая в ее отношении права владения, пользования и распоряжения. Рекомендации FATF предписывают национальным органам финансовой разведки устанавливать нормы, касающиеся регулирования криптовалюты, в том числе, идентификация при совершении операций и исследование ее происхождения. С этим тесно связана задача оценки заемщиков, берущих кредиты под залог криптовалюты. Неверная оценка заемщика может сделать, в том числе, и добросовестного кредитора стороной нелегальной сделки. Здесь задача финансово-кредитных институтов оказывается тесно увязанной с более общей проблемой выявления в сетях, использующих криптовалюты, участников, ведущих подозрительную деятельность. Получение кредита под залог криптовалюты является одной из форм конвертации криптовалютных активов в фиатные деньги, и очевидным образом может быть использовано в преступных целях. Тем самым построение моделей кредитного скоринга для сетей распределенных реестров оказывается включенным в общую проблематику исследований, направленных на оценку криптовалютных активов на подозрительность происхождения и наличие признаков отмывочной деятельности.

В последние годы, сопровождающиеся взлетом цены на биткоин, получили распространение кредиты под залог криптовалюты, см. [1]. Свидетельством этого является рост числа запросов относительно таких кредитов, демонстрирующий определенную корреляцию с курсом биткоина (по данным Google Trends (рис.1)).

Естественной формой кредитования под залог криптовалюты становится одноранговое (P2P) кредитование [2]. Одноранговое кредитование позволяет участникам сети получать кредиты и выдавать займы, минуя посредников, в том числе финансовые институты. В децентрализованной системе P2P-кредитования, использующей технологию распределенных реестров,

1 Гарин Леонид Андреевич, главный аналитик данных, ПАО Банк ЗЕНИТ, г. Москва, Россия. E-mail: l.garin@zenit.ru

2 Гисин Владимир Борисович, кандидат физико-математических наук, профессор, профессор Департамента математики Финансового университета при Правительстве Российской Федерации, г. Москва, Россия. E-mail: vgisin@fa.ru

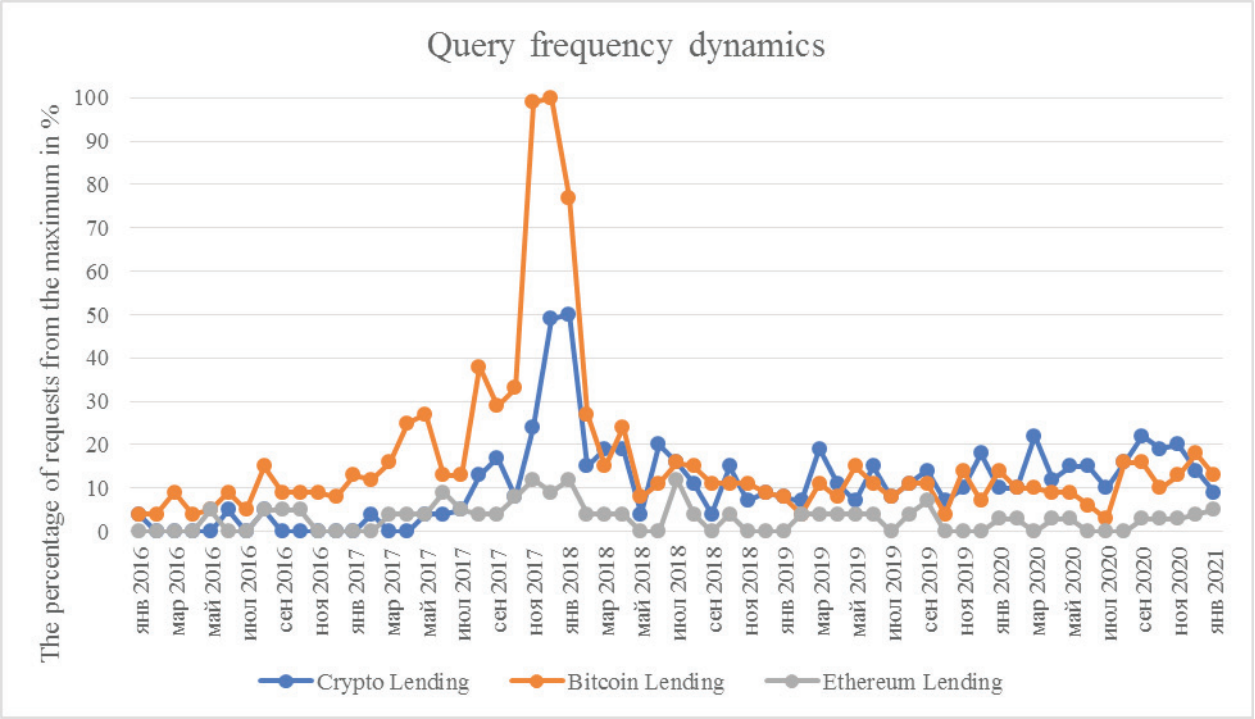

Рис. 1. Динамика интереса пользователей к кредитам, обеспеченным криптовалютами. Источник - Google Trends, расчет авторов

кредитование может осуществляться между любыми заинтересованными пользователями без участия какого-либо финансового учреждения и обеспечивать лучшие условия для заемщиков и кредиторов.

Применение смарт-контрактов, поддерживаемых децентрализованной сетью, вообще говоря, ускоряет процесс кредитования, снижает накладные расходы и дает определенные гарантии относительно безопасности на основе оценки рисков. В то же время специфика цифровой погруженности вызывает ряд проблем. Например, проблемой является конвертация цифровых токенов для несвязанных сетей распределенных реестров. Еще одной проблемой является оценка (скоринг) заемщиков. В экстремально-цифровом варианте они могут быть представлены лишь своим цифровым профилем (или, вообще, кошельком, как это происходит в сетях блокчейн, не требующих разрешений). Эти две проблемы в некотором смысле полярны.

Решение первой проблемы лежит в области технологий. Ее решение могут обеспечить соответствующим образом разработанные платформы и протоколы взаимодействия. Например, так называемая технология атомарных свопов. Эта технология позволяет проводить кредитование под залог криптовалюты и обеспечивает ликвидность залога в случае дефолта заемщика [3].

Для решения второй проблемы требуется изучить и оценить цифровую версию кредитной истории заемщика. Причем это должно быть сделано в среде, где анонимность и деиндивидуализация считаются одними из главных преимуществ и обеспечиваются сетевыми протоколами. В своей постановке эта проблема, безусловно, родственна проблеме традиционного скоринга, однако специфика криптовалют приводит к появлению целого ряда задач, характерных именно для этого вида залогового кредитования. Решение этих задач позволяет снизить риски, связанные с кредитованием под залог криптовалюты. В настоящей работе приводится описание некоторых из этих задач и подход к их решению на основе методов теории машинного обучения и искусственного интеллекта.

Для получения сравнительно небольших сумм при P2P-кредитовании обеспечение, как правило, не требуется. Для более крупных сумм, может потребоваться залог. В P2P-кредитовании даже при отсутствии залога средняя процентная ставка, как правило, ниже, чем в среднем по стране (для стран БРИКС), но с более короткими сроками погашения, чем при традиционном кредитовании. Используя платформу блокчейн, заемщик может ввести информацию о себе (о проекте), загрузить поддающиеся проверке документы и запросить сумму кредита. Платформы, обеспечивающие P2P-кредитование, используют несколько источников

проверки, чтобы присвоить этому запросу оценку качества, форму кредитного рейтинга. Кредиторы предлагают ставки по процентам и погашению. Как только запрошенная сумма собрана и принята, в игру вступает смарт-контракт. Минимальная процентная ставка платформ, выдающих кредиты под залог криптовалюты, колеблется в промежутке от 1% до 6%. Максимальные процентные ставки для кредиторов – от 9,6% до 21,49%, максимальный LTV (Lifetime Value – прибыль от отношений с клиентом за весь период) платформ – от 50% до 83%. Сроки займов, как правило, находятся в районе полугода, но возможны и займы с более короткими (до 1 месяца) и более длинными (до трех лет) сроками.

Схематично платформу кредитования под залог криптовалюты можно описать следующим образом. Прежде всего, клиент должен зарегистрироваться. Затем он формирует заявку на кредит. Платформа в ответ указывает условия кредита, в частности, объем залога в криптовалюте, и предлагает клиенту предъявить доказательство владения криптовалютой в нужном объеме. После предъявления доказательств платформа производит оценку залога. В эту оценку помимо ликвидности и рыночной стоимости входит оценка вовлеченности в запрещенные виды деятельности (мошенничество, отмывание доходов, финансирование терроризма). На основе проверки принимается решение о выдаче кредита. В случае положительного решения выдача кредита и прием залога оформляются соответствующим смарт-контрактом. Риски кредитора платформа принимает на себя, покрывая убытки кредитора в случае их появления. Заметим, что некоторыми платформами P2P-кредитования предусмотрены пользователи-коллекторы, действующие в цифровом пространстве.

Чтобы снизить риск, кредиторам требуется иметь более подробную информацию о заемщике, как пользователе цифровой системы, чем ту, которая представлена в его профиле. Подобная информация может быть получена в результате анализа транзакций заемщика – своеобразной замены кредитной истории. Достаточно полная информация о транзакциях содержится в соответствующих базах данных. Проблема в том, как по транзакциям выявить «подозрительного» заемщика. Та же проблема может возникнуть и у более традиционных институтов кредитования, если в качестве залога выступает криптовалюта.

В настоящей работе приводится описание алгоритма выявления преступных активностей цифрового клиента в сети блокчейн. Исследование находится в русле разработки подозрительных шаблонов поведения, формируемых и анализируемых в течение уже более, чем десяти лет. В частности, рассматривается задача выявления подозрительных клиентов в сети Биткоин. Основы формирования шаблонов были созданы спецслужбами [4], однако применение предложенных ими методов создателями платформ P2P-кредитования, не обеспечивает достаточной эффективности, см. [5]. Сложность анализа транзакций в сети Биткоин объективна и обусловлена отсутствием привязки транзакций к владельцу адреса. Таким образом, метку «подозрительности» должны получать не клиенты сети, а адреса. В работе предложен подход на основе интеллектуального анализа данных. Экспериментальное изучение различных методов анализа данных показало, что наилучший результат дает градиентный бустинг. Качество, сопоставимое и лучшее, чем в наиболее продвинутых исследованиях, достигается здесь при использовании меньшего числа признаков и меньшей глубины агрегации данных, что свидетельствует о надежности и достоверности результатов, ср. [6]. Результатом является оценка меры принадлежности узла к числу отмывочных сервисов, сервисов, связанных с азартными играми, майнинговыми пулами, обменниками криптовалют. Усредненная метрика качества F1-Score 0.95.

Полученные результаты свидетельствуют о необходимости дальнейшего совершенствования методов скоринга узлов сетей распределенных реестров, основанных на интеллектуальном анализе данных.

Основная цель этой статьи – дать описание применения современных методов машинного обучения для решения проблемы классификации вредоносных биткойн-адресов на основе данных о транзакциях и продемонстрировать результаты применения этих методов. Основным источником данных для анализа служит набор данных о транзакциях Elliptic Data Set – самый большой массив данных (информация о более, чем 200 тыс. транзакций) для идентификации транзакций, связанных с отмыванием денег с использованием криптовалют. В разделе 1 дается общая постановка проблемы. В разделе 2 обсуждается проблема разметки биткоин-адресов и сбора данных о транзакциях. В разделе 3 приведено описание набора данных Elliptic Data Set и обзор результатов, полученных в моделях на основе данных из Elliptic Data Set. В разделе 4 приведено описание оригинальной технологии подготовки и обработки данных.

1. Криптовалюты и нелегальная финансовая деятельность

Появление криптовалюты Биткоин ознаменовало новую эру криминальных транзакций. Сама по себе

криптовалюта обеспечивает некоторый уровень анонимности, позволяя своим пользователям создавать неограниченное количество кошельков с псевдонимными адресами, что затрудняет идентификацию реального пользователя. Это используется преступниками с целью совершения незаконных операций, связанных с отмыванием денег, финансированием терроризма, незаконным оборотом наркотиков и т.д. В то же время Биткоин хранит и предоставляет публично информацию обо всех совершенных транзакциях, что открывает возможности для выявления подозрительных моделей поведения в этой сети с помощью интеллектуального анализа данных. Опубликованный набор данных Elliptic содержит разметку адресов в сети Биткоин и используется исследователями для построения моделей машинного обучения на основе особенностей транзакций. Это позволяет разрабатывать методы контролируемого обучения для решения описанной задачи, и анализировать их эффективность.

Криптовалюта – это «электронная наличность», основанная на электронной записи в открытом децентрализованном распределенном реестре. Криптографические алгоритмы с одной стороны защищают информацию реестра от несанкционированного изменения, с другой – обеспечивают ее доступность участникам – владельцам адресов (узлов, нодов). Принципы работы самой популярной криптовалюты Биткоина были разработаны и представлены в 2008 году [7]. Цель создания биткоина – снизить затраты за счет устранения посредников при совершении транзакций. Подтверждение транзакций осуществляется с использованием алгоритмов консенсуса – согласия сообщества пользователей на запись блока транзакций в реестр.

Биткоин часто используется для подозрительной или преступной деятельности: мошенничества, вымогательства, торговли на черном рынке, а также отмывания денег и финансирования терроризма (ОД/ФТ). Теоретически Биткоин обеспечивает возможность полностью анонимных транзакций. Для полной анонимизации требуется создавать новый адрес для каждой новой операции и не совершать транзакции более чем с одним отправителем. Однако пользователи, в том числе киберпреступники, часто игнорируют эти правила. Таким образом, становится возможным связывать группы адресов, выявлять подозрительные адреса и транзакции с помощью специальных методов, включая анализ больших данных [8].

Распределенный реестр транзакций может быть представлен в виде ориентированного графа, вершинами которого являются адреса кошельков, а ребрами – транзакции. Кроме того, можно присвоить ребрам некоторую дополнительную информацию: сумму транзакции, временную метку, комиссию майнера за включение в блок и т.д. Таким образом, для каждого узла можно собрать его характеристики как вершины графа и использовать их в качестве предикторов в моделях машинного обучения.

В этих условиях перед органами финансового мониторинга многих стран стоит задача выявления описанной выше деятельности. Международная Группа разработки финансовых мер борьбы с отмыванием денег (ФАТФ) регулярно обновляет рекомендации по установлению правового режима идентификации клиентов при операциях с криптовалютами, в частности, требуя соблюдения, так называемого, принципа «Знай своего клиента» (KYC) [9]. Однако предлагаемая юридическая ответственность, связанная с необходимостью идентификации клиента, по-видимому, не в полной мере эффективна из-за экстерриториального принципа Биткоина. В связи с этим становится актуальным изучение существующего опыта применения методов машинного обучения для классификации подозрительных биткоин-адресов на основе характеристик транзакций.

Анализ статей в Google Scholar показывает, что выявление подозрительных закономерностей в сетевых структурах изучалось исследователями из разных стран на протяжении многих лет. Методов выявления сетевых аномалий нашли применение во многих областях: обнаружение мошенничества с кредитными картами [10], кибербезопасность [11], социальные сети [12] и др. Согласно Глобальному обзору экономической преступности за 2016 год [13], операции по отмыванию денег во всем мире оцениваются примерно в 1-2 триллиона долларов США в год. В том же исследовании говорится, что только 50% случаев отмывания денег или финансирования терроризма были выявлены с помощью систем предупреждения. В [14] утверждается, что алгоритмы обнаружения мошенничества (модели борьбы с мошенничеством), используемые в банках, неприменимы к специфике криптовалют. Однако в [15] представлены методы деанонимизации пользователей Биткоина, что позволяет запрашивать информацию, непосредственно характеризующую человека, и, соответственно, применять модели, использующие предикторы, основанные на «традиционных» социально-демографических данных.

Одним из классических методов, используемых в разведке правоохранительных органов, является

анализ сетевых подключений для выявления отмывания денег [16]. Распределение связей между узлами, их структура и связность в отдельных подграфах характеризуют сообщества – таким образом, при присвоении ярлыка «незаконный» нескольким членам сообщества он может быть (при определенных предположениях) обобщен на всех участников.

До недавнего времени сеть Биткоин оставалась привлекательной для киберпреступников, поскольку правоохранительные органы сталкивались с трудностями при обнаружении подозрительной активности, идентификации реальных пользователей и получении записей транзакций [17]. Сегодня как правоохранительные органы, так и организации, которым необходимо отслеживать подозрительную активность, разрабатывают и внедряют методологии выявления подозрительных поведенческих моделей на основе характеристик, отслеживаемых в публичной сети Биткоин (см., например [18], где описан подход ФБР США).

2. Сбор меток и данных о транзакциях

Основная проблема с применением методов классификации машинного обучения заключается в том, что адреса в сети Биткоин в их «сыром» виде не размечены в соответствии с принадлежностью к определенной категории. В блокчейне Биткоина нет явной ссылки на владельца адреса. Таким образом, адресные метки, идентифицирующие «честные» или «подозрительные» узлы, должны быть присвоены исследователем или третьей стороной на основе опыта или специальных методов [19].

Имеются сервисы, предоставляющие общедоступные базы данных помеченных адресов.

WalletExplorer. Сервис с помощью автоматического поиска в Интернете идентифицирует адреса и подключает их к кошелькам. Некоторые кошельки помечены тегами, соответствующими сервисам, на которых они были найдены. Есть ярлыки азартных игр, криптовалютных бирж, магазинов даркнета, мошеннических сайтов финансовых пирамид, микширующих сервисов.

BitcoinAbuse. Открытая база данных с отчетами о вредоносном поведении, касающемся конкретных кошельков. Чаще всего пользователи помечают адреса как спам, реже пометки связаны с вымогательством, мошенничеством и другими преступлениями. Хотя информацию от одного конкретного пользователя трудно проверить, можно отфильтровать адреса, по которым были совершены транзакции и/или было подано несколько жалоб.

Эти сервисы имеют интерфейс, который позволяет автоматически собирать статистическую информацию о «честных» и «подозрительных» агентах.

Сбор информации об адресах транзакциях может осуществляться несколькими способами. Первый способ – загрузить полный блокчейн и собрать информацию обо всех транзакциях с помощью парсера. Этот метод требует значительного дискового пространства и навыков синтаксического анализа, но после подготовки позволяет быстро собрать всю информацию, включая данные транзакций, относительно конкретного адреса. Второй способ – использовать внешних поставщиков данных и запрашивать информацию с помощью API (интерфейса программирования приложений). Однако, поскольку количество транзакций по одному адресу может достигать внушительных значений, поставщики API обычно налагают ограничения на число запросов в единицу времени и на число транзакций в одном запросе.

3. Моделирование на основе Elliptic dataset

Набор данных Elliptic dataset [20], содержащий информацию о 203 769 транзакциях был подготовлен частной компанией Elliptic, специализирующейся на защите криптовалютных экосистем от преступной деятельности. Около 2% транзакций помечены как «подозрительные», совершенные подозрительной организацией, такой как мошенники, вредоносные программы, террористические организации, программы-вымогатели, схемы Понци. Около 21% от общего числа транзакций принадлежат легальным пользователям – биржам, поставщикам кошельков, майнерам, легальным сервисам и т.д. Остальные транзакции неразмечены.

Для каждой транзакции известно астрономическое время ее включения в блок и подтверждения сетью Биткоин. В рассматриваемом наборе данных используются 49 временных интервалов, равномерно распределенных на промежутке около двух недель (конкретные даты не указаны). Каждый временной интервал сравнивается с определенным списком связанных транзакций, которые появились в блокчейне менее, чем за 3 часа. Следует отметить, что в наборе нет ребер, соединяющих разные временные шаги.

Для каждого узла набор данных содержит 166 признаков, основанных на информации о самом узле. Следует отметить, что набор содержит адреса и признаки в легендированной форме. Перед публикацией в отношении предикторов были предприняты некоторые действия по обфускации (засекречиванию) ин-

Таблица 1

Показатели оценки моделей бинарной классификации

Алгоритм	Precision	Recall	F1-Score
Логит регрессия	0.404	0.593	0.481
Случайный лес	0.956	0.670	0.788
Многослойный персептрон	0.694	0.617	0.653
Сверточная нейронная сеть	0.812	0.512	0.628

Таблица 2

Показатели оценки моделей с учетом динамической структуры данных

Алгоритм	Precision	Recall	F1-Score
XGBoost	0.875	0.742	0.803
LightGBM	0.863	0.741	0.797
Random Forest	0.875	0.723	0.792
AXGB – original modification (timestep ≥ 5)	0.813	0.680	0.740

формации: предикторы стандартизированы (среднее значение 0, стандартное отклонение 1), а их имена скрыты. В [20] говорится, что первая группа из 94 предикторов представляет локальную информацию о транзакции, включая временной шаг, количество входов и выходов, плату майнера, сумму входа и некоторые агрегированные значения: средняя сумма и количество входов и выходов транзакции. Вторая группа из 72 предикторов содержит агрегированную информацию о «соседних» транзакциях – на один шаг вперед/назад от основного узла. Агрегирование проводилось путем вычисления минимального, максимального и стандартного отклонения признаков, аналогичных признакам первой группы. В то время как сами данные могут быть использованы для построения классификаторов, обфускация предикторов оставляет открытым вопрос о том, как эти данные фактически используются для идентификации подозрительных адресов в реальной среде.

Проблема, рассматриваемая в работе, может быть сведена к задаче бинарной классификации. В целом задачу можно сформулировать следующим образом.

Имеется множество объектов X (потенциально бесконечное) и множество меток $Y = \{y_1, \ldots, y_s\}$. Задана конечная обучающая выборка $S \subset X \times Y$, являющаяся функциональным отношение на своей области определения. Требуется построить алгоритм вычисления функции $a = X \rightarrow Y$, график которой содержит S.

В [20] проводится бинарная классификация (с $s=2$) для прогнозирования незаконных транзакций с использованием логит регрессии, случайного леса, многослойного персептрона и сверточной нейронной сети. Данные были разделены на обучающие и тестовые подмножества с соотношением 70:30. Результаты по тестовому подмножеству показывают преимущество случайного леса (см. таблицу 1). Также предложена концепция инструмента для визуализации производительности модели при перемещении по графу транзакций.

В дальнейшем была сделана попытка улучшить результат за счет настройки гиперпараметров и борьбы с дисбалансом классов набора данных [21].

В [22], также используя Elliptic dataset, авторы предлагают модификацию алгоритма XGBoost, учитывающую динамическую структуру данных. Результаты (см. Таблицу 2) превосходят предыдущее решение, как благодаря новому алгоритму обучения, так и благодаря применяемому методу перебалансировки классов NCL-SMOTE для решения проблемы низкого процента незаконных меток в исходном наборе данных.

Борясь с дисбалансом классов, с недостаточной разметкой, исследователи пытаются «выжать максимум» из данных, предоставляемых Elliptic Dataset. Интересным примером, является работа [23], в которой предлагается использовать активное обучение. Предполагается, что, используя активное обучение, алгоритм способен достичь сопоставимого качества при меньшем количестве помеченных данных по сравнению с классическим обучением под наблюдением – эта предпосылка была экспериментально подтверждена в рассматриваемой работе.

Подводя итоги, можно сделать вывод, что следующие характеристики являются значимыми в моделях классификации подозрительных адресов:

- сумма транзакций (средняя, минимальная, максимальная, медианная, размах);
- частота транзакций;
- количество уникальных кошельков, указанных в отправителях/получателях по транзакции с данным кошельком;
- характеристики адресов, с которыми были транзакции.

Таким образом, задача выявления подозрительной активности в сети Биткоин решается методами интеллектуального анализа данных на основе графовых характеристик узлов (адресов) и связей (транзакций).

4. Технология выявления подозрительных адресов

В проведенном исследовании использован размеченный набор Bitcoin-адресов[3], опубликованный на сайте Harvard Dataverse – открытом репозитории Гарвардского университета. В наборе представлена информация о блоках с высотой от 520890 (временная метка 2018-05-02 16:29) до 520910 (2018-05-02 20:06). Путем сбора информации с WalletExplorer, публичных форумов, веб-сайтов конкретных сервисов размечено 8808 адресов для следующих типов узлов (в скобках указано число адресов):

- майнеры (4030)) – пользователи, которые занимаются подтверждением блоков транзакций;
- майнинговые пулы (89) – объединения майнеров;
- обменники валют (1666) – площадки, предоставляющие возможность купить или продать Bitcoin в обмен на фиатные деньги или другую криптовалюту;
- миксинговые сервисы (800) – площадки, предоставляющие услуги по скрытию транзакций, связанных, как правило, с нелегальной торговлей;
- площадки для приема ставок и азартных игр (911);
- прочие сервисы (1312) – площадки, принимающие Bitcoin в качестве оплаты.

Для сбора информации о транзакциях, связанных с этими адресами, использовались два сервиса – обозревателя блокчейна: BitAps.com, Blockchain.Info. Сначала с помощью BitAps.com выгружались хеши последних 250 транзакций для каждого адреса. Затем для каждой выгруженной на предыдущем шаге транзакции, с сайта Blockchain.Info была получена детальная информация об отправителях, получателях, суммах и прочие данные. Всего было собрано 578809 транзакций.

Собранная информация о транзакциях адреса позволяет построить на ее основе предикторы и далее обучить модели классификации адресов.

Преобразование данных осуществлено на двух уровнях:

- каждая транзакция преобразуется в одномерный набор признаков фиксированной длины;
- для каждого адреса собирается агрегированная информация по транзакциям, в которых адрес принимал участие в качестве отправителя или получателя, и рассчитываются производные предикторы.

Для предикторов оценены распределения их значений для каждого типа адресов. Укажем некоторые характерные черты распределений

Для миксинговых сервисов характерна повышенная комиссия – причем, для майнинговых пулов – пониженная. Майнинговые пулы совершают много coinbase-транзакций, поэтому плотность среднего числа отправителей по адресу сильно смещена влево. Напротив, обменники валют имеют среднее число отправителей больше остальных. Острый пик распределения среднего числа исходящих адресов в транзакциях для миксинговых сервисов говорит о сильной концентрации количества адресов-получателей, медиана составляет два адреса. Видно, что у адресов, принадлежащих обменникам, число как отправителей, так и получателей смещается вправо. Входящие суммы для адресов, принадлежащих миксинговым сервисам, сильно сконцентрированы: медианное значение – 6.69 BTC. Это подтверждает предположение о том, что частое повторение транзакций с одной и той же суммой является маркером подозрительности.

В дальнейшем для практического моделирования воспользуемся свободно распространяемым пакетом машинного обучения scikit-learn для языка программирования Python 3. Во всех моделях использованы гиперпараметры, найденные с помощью случайного поиска по решетке (Random Grid Search из пакета sklearn). Выборка была случайным образом разделена на обучающую и контролирующую для каждого алгоритма в соотношении 80:20.

Испытания проводились на следующих моделях обучения: логистическая регрессия; дерево решений; случайный лес; градиентный бустинг.

3 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KEWU0N

Таблица 3

Метрики для градиентного бустинга

Тип узла (кол-во)	Precision	Recall	F1-Score
майнеры (18)	0,92	0,97	0,94
майнинговые пулы (18)	1,00	1,00	1,00
обменники (338)	0,92	0,87	0,89
миксинговые сервисы (161)	1,00	0,97	0,98
азартные игры (166)	0,99	0,95	0,97
прочие сервисы (259)	0,97	0,91	0,94

Наилучшие результаты получены при градиентном бустинге (табл. 3).

Отметим, что удалось достичь качества, лучшего, чем в [24], при использовании меньшего числа признаков (38 признаков) и меньшей глубины агрегации данных.

Заключение

Задача обнаружения подозрительной активности в сети Биткоин может быть решена с достаточно высокой точностью с помощью методов машинного обучения, контролируемых на основе графовых характеристик узлов (адресов) и соединений (транзакций). Наилучшие результаты в определении подозрительных узлов сети Биткоин показывают метод случайного леса и метод градиентного бустинга. Важную роль в построении эффективных систем машинного обучения применительно к рассматриваемой проблеме играет методика выбора гиперпараметров, перебалансировки набора данных и обучения с использованием фреймворка активного обучения. Описанные в работе методы выявления подозрительных узлов могут служить основой практически значимой технологии и быть внедрены в реальные бизнес-процессы.

Литература

1. Wang R., Lin Z., Luo H. Blockchain, bank credit and SME financing //Quality & Quantity. 2019. Т. 53. №. 3. С. 1127-1140.
2. Arvanitis S. P2P lending review, analysis and overview of Lendoit blockchain platform //International Journal of Open Information Technologies. 2019. Т. 7. №. 2. С. 94-98.
3. Black M., Liu T., Cai T. Atomic loans: Cryptocurrency debt instruments //arXiv preprint arXiv:1901.05117. 2019. 1-13.
4. Federal Bureau of Investigation. Bitcoin Virtual Currency: Unique Features Present Distinct Challenges for Deterring Illicit Activity. 2012. Available at: http://www.wired.com/images_blogs/threatlevel/2012/05/Bitcoin-FBI.pdf. Последнее обращение 27.05.2022
5. Berg A. The identity, fungibility and anonymity of money //Economic Papers: A journal of applied economics and policy. 2020. Т. 39. №. 2. С. 104-117.
6. Michalski R., Dziubałtowska D., Macek P. Revealing the character of nodes in a blockchain with supervised learning //Ieee Access. 2020. Т. 8. С. 109639-109647.
7. Nakamoto, Satoshi. Bitcoin: A peer-to-peer electronic cash system. Manubot, 2019.
8. Zheng B., Zhu, L., Shen M., Du, X., Guizani, M. Identifying the vulnerabilities of bitcoin anonymous mechanism based on address clustering //Science China Information Sciences. – 2020. – Т. 63. – №. 3. – С. 1-15.
9. Alkadri S. Defining and regulating cryptocurrency: fake internet money or legitimate medium of exchange //Duke L. & Tech. Rev. – 2018. – Т. 17. – С. 71-98.
10. Belle R. V., Mitrović S., Weerdt J. D. Representation learning in graphs for credit card fraud detection //Workshop on Mining Data for Financial Applications. – Springer, Cham, 2019. – С. 32-46.
11. Böhm F., Menges F., Pernul G. Graph-based visual analytics for cyber threat intelligence //Cybersecurity. – 2018. – Т. 1. – №. 1. – С. 1-19.
12. Akhtar N., Ahamad M. V. Graph tools for social network analysis //Research Anthology on Digital Transformation, Organizational Change, and the Impact of Remote Work. – IGI Global, 2021. – С. 485-500.
13. Global Economic Crime Survey 2016. PwC 2016. С. 1-56 – URL: https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf (Дата обращения 27.05.2022).
14. Marcin S. I. Bitcoin Live: scalable system for detecting bitcoin network behaviors in real time / ACADEMIA. 2015. С. 1-12
15. ShenTu Q. C., Yu J. P. Research on Anonymization and De-anonymization in the Bitcoin System //arXiv preprint arXiv:1510.07782. – 2015. С.1-14.
16. Bolton R. J., Hand D. J. Statistical fraud detection: A review //Quality control and applied statistics. – 2004. – Т. 49. – №. 3. – С. 313-314.
17. Meiklejohn S., Pomarole M., Jordan G., Levchenko K., McCoy D., Voelker G. M., Savage S. A fistful of bitcoins: characterizing payments among men with no names //Proceedings of the 2013 conference on Internet measurement conference. – 2013. – С. 127-140.

18. Federal Bureau of Investigation. Bitcoin Virtual Currency: Unique Features Present Distinct Challenges for Deterring Illicit Activity. [Электронный ресурс] – 2012. Режим доступа: http://www.wired.com/images_blogs/threatlevel/2012/05/Bitcoin-FBI.pdf. Дата доступа 27.05.2022
19. Wang M., Ichijo H., Xiao B. Cryptocurrency address clustering and labeling //arXiv preprint arXiv:2003.13399. – 2020. С.1-7.
20. Weber M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics //arXiv preprint arXiv:1908.02591. – 2019. С.1-7.
21. Фельдман Е.В., Ручай А.Н., Матвеева В.К., Самсонова В.Д. Модель выявления аномальных транзакций биткоинов на основе машинного обучения // Челябинский физико-математический журнал. 2021. №1. С. 119-132.
22. Vassallo D., Vella V., Ellul J. Application of gradient boosting algorithms for anti-money laundering in cryptocurrencies //SN Computer Science. – 2021. – Т. 2. – №. 3. – С. 1-15.
23. Lorenz J., Silva M. I., Aparício D., Ascensão J. T., Bizarro P. Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity //Proceedings of the First ACM International Conference on AI in Finance. – 2020. – С. 1-8.
24. Michalski R., Dziubałtowska D., Macek P. Revealing the character of nodes in a blockchain with supervised learning //IEEE Access. 2020. Т. 8. С. 109639-109647.

IDENTIFICATION OF SUSPICIOUS BITCOIN NETWORK NODES BY BIG DATA ANALYSIS METHODS

Garin L.A[4], Gisin V.B.[5]

Purpose: *to analyze existing machine learning models that allow identifying suspicious addresses of the Bitcoin network, to develop a modern effective scoring model for identifying suspicious addresses.*

Methods: *collecting and analyzing data on addresses and transactions of the Bitcoin network, identification of patterns of illicit activity of addresses, developing and experimental verification of machine learning models aimed at identifying suspicious addresses of the Bitcoin network using related transactions.*

Practical relevance: *the analysis of common data sets and machine learning models related to the identification of suspicious Bitcoin network addresses was carried out, data on transactions related to a representative set of addresses was collected. Machine learning models have been built to identify suspicious addresses based on the collected information. Experimental approbation of the models was carried out. It is established that the best result is obtained by a model using gradient boosting. This model demonstrates more efficient operation compared to existing analogues.*

Keywords: *distributed ledger, blockchain, transaction, scoring, machine learning, cryptocurrency, illicit activity.*

References

1. Wang R., Lin Z., Luo H. Blockchain, bank credit and SME financing //Quality & Quantity. 2019. Т. 53. №. 3. S. 1127-1140.
2. Arvanitis S. P2P lending review, analysis and overview of Lendoit blockchain platform //International Journal of Open Information Technologies. 2019. Т. 7. №. 2. S. 94-98.
3. Black M., Liu T., Cai T. Atomic loans: Cryptocurrency debt instruments //arXiv preprint arXiv:1901.05117. 2019. 1-13.
4. Federal Bureau of Investigation. Bitcoin Virtual Currency: Unique Features Present Distinct Challenges for Deterring Illicit Activity. 2012. Available at: http://www.wired.com/images_blogs/threatlevel/2012/05/Bitcoin-FBI.pdf. Data obrashhenija 27.05.2022
5. Berg A. The identity, fungibility and anonymity of money //Economic Papers: A journal of applied economics and policy. 2020. Т. 39. №. 2. S. 104-117.
6. Michalski R., Dziubałtowska D., Macek P. Revealing the character of nodes in a blockchain with supervised learning //Ieee Access. 2020. Т. 8. S. 109639-109647.
7. Nakamoto, Satoshi. Bitcoin: A peer-to-peer electronic cash system. Manubot, 2019.
8. Zheng B., Zhu, L., Shen M., Du, X., Guizani, M. Identifying the vulnerabilities of bitcoin anonymous mechanism based on address clustering //Science China Information Sciences. – 2020. – Т. 63. – №. 3. – S. 1-15.

4 Leonid A. Garin, Chief Data Analyst, PJSC Bank ZENIT, Moscow, Russia. E-mail: l.garin@zenit.ru

5 Vladimir B. Gisin, Ph.D., Professor, Professor of the Department of Mathematics, Finance University under the Government of the Russian Federation, Moscow, Russia. E-mail: vgisin@fa.ru

9. Alkadri S. Defining and regulating cryptocurrency: fake internet money or legitimate medium of exchange //Duke L. & Tech. Rev. – 2018. – Т. 17. – S. 71-98.
10. Belle R. V., Mitrović S., Weerdt J. D. Representation learning in graphs for credit card fraud detection //Workshop on Mining Data for Financial Applications. – Springer, Cham, 2019. – S. 32-46.
11. Böhm F., Menges F., Pernul G. Graph-based visual analytics for cyber threat intelligence //Cybersecurity. – 2018. – Т. 1. – №. 1. – S. 1-19.
12. Akhtar N., Ahamad M. V. Graph tools for social network analysis //Research Anthology on Digital Transformation, Organizational Change, and the Impact of Remote Work. – IGI Global, 2021. – S. 485-500.
13. Global Economic Crime Survey 2016. PwC 2016. C. 1-56 – URL: https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf Poslednee obrashhenie 27.05.2022.
14. Marcin S. I. Bitcoin Live: scalable system for detecting bitcoin network behaviors in real time / ACADEMIA. 2015. S. 1-12
15. ShenTu Q. C., Yu J. P. Research on Anonymization and De-anonymization in the Bitcoin System // arXiv preprint arXiv:1510.07782. – 2015. C.1-14.
16. Bolton R. J., Hand D. J. Statistical fraud detection: A review //Quality control and applied statistics. – 2004. – Т. 49. – №. 3. – S. 313-314.
17. Meiklejohn S., Pomarole M., Jordan G., Levchenko K., McCoy D., Voelker G. M., Savage S. A fistful of bitcoins: characterizing payments among men with no names // Proceedings of the 2013 conference on Internet measurement conference. – 2013. – S. 127-140.
18. Federal Bureau of Investigation. Bitcoin Virtual Currency: Unique Features Present Distinct Challenges for Deterring Illicit Activity. 2012. Available at: http://www.wired.com/images_blogs/threatlevel/2012/05/Bitcoin-FBI.pdf. (Accessed 27.05.2022)
19. Wang M., Ichijo H., Xiao B. Cryptocurrency address clustering and labeling //arXiv preprint arXiv:2003.13399. – 2020. C.1-7.
20. Weber M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics //arXiv preprint arXiv:1908.02591. – 2019. C.1-7.
21. Fel'dman E.V., Ruchaj A.N., Matveeva V.K., Samsonova V.D. Model' vyjavlenija anomal'nyh tranzakcij bitkoinov na osnove mashinnogo obuchenija // Cheljabinskij fiziko-matematicheskij zhurnal. 2021. №1. S. 119-132.
22. Vassallo D., Vella V., Ellul J. Application of gradient boosting algorithms for anti-money laundering in cryptocurrencies //SN Computer Science. – 2021. – Т. 2. – №. 3. – S. 1-15.
23. Lorenz J., Silva M. I., Aparício D., Ascensão J. T., Bizarro P. Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity //Proceedings of the First ACM International Conference on AI in Finance. – 2020. – S. 1-8.
24. Michalski R., Dziubałtowska D., Macek P. Revealing the character of nodes in a blockchain with supervised learning //IEEE Access. 2020. T. 8. S. 109639-109647.