A work project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

1. Segmenting Economic Agents on the XRP Ledger: A Heuristic-Driven Approach

LUISA REICHENBACH

Work project carried out under the supervision of:

Prof. Leid Zejnilovic

17/12/2024

Abstract

XRP is among the top 5 most prevalent cryptocurrencies as of 2024. However, the pseudonymous nature of XRP Ledger transactions complicates the understanding of economic activities and regulatory oversight of fraud within the network. This study uses descriptive statistics to develop heuristics for categorizing distinct types of economic agent activities.

Keywords

Behavioral Analysis, Blockchain, Heuristics, Ledgerlytics, Machine Learning, Ripple, Segmentation, XRP Ledger

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1 Introduction

Since its introduction in 2009, Bitcoin has significantly influenced the global monetary system, leading to the emergence of almost ten thousand competing cryptocurrencies on the market (Statista, n.d.; Alahmad et al. 2023). The XRP token, as one of these currencies, was introduced by Ripple Inc. in 2012 and is widely recognized as one of the most prominent alternatives to Bitcoin. XRP ranks in the top 5 digital currencies worldwide with a market capitalization of over 100 billion USD as of November 2024 (Ahmadova and Erek 2022; CoinMarketCap 2024).

The XRP network mainly aims to serve banks and financial institutions by providing a decentralized alternative to traditional cross-border payment systems such as the Society for Worldwide Interbank Financial Telecommunication (SWIFT), offering superior transaction speed and reduced costs for currency exchanges (Sergeenkov 2024). Additionally, the network supports key applications, including tokenization, central bank digital currencies, and stablecoins, as well as Non-Fungible Tokens (Ripple, n.d.-b).

While the XRP Ledger, as the decentralized ledger technology behind its token, operates as a publicly accessible database, the information recorded on-chain alone is insufficient to trace users back to their real-world counterparts. In blockchain, participants transact assets using unique alphanumeric addresses. These addresses lack a direct connection to real-world identities or individual agents, a trait known as pseudonymity (Hellwig, Karlic, and Huchzermeier 2020).

This characteristic inherently results in a lack of transparency and uncertainty regarding the types of economic activities in blockchain networks such as the XRP Ledger. The limited traceability of transactions also raises concerns about the potential misuse of the technology for illicit purposes. Illegal crypto activities such as scams, stolen funds, and illicit trade accounted for a total estimate of 39.6 billion USD in 2022 (Chainalysis 2024). As a result, stakeholders – including governments, industry leaders, and financial institutions – require a shared

understanding of the participants within blockchain ecosystems to foster sustainable growth and address the vulnerabilities inherent to the networks (Chainalysis, n.d.).

Addressing these challenges, this study seeks to analyze agents on the XRP Ledger to characterize the economic activities within the ecosystem and aims to detect accounts involved in selected fraud schemes. The main objectives of the study is to explore the applicability of account-based heuristics in characterizing different economic agent categories and their behavior on the XRP Ledger.

To the best of our knowledge, no similar research has been conducted on XRP, leading to uncertainties about the applicability of existing methods to gain insights into the network. While modeling techniques, such as machine learning models have been applied with similar aims to Bitcoin and Ethereum, their relevance and effectiveness on other ecosystems remains unclear, as ledger structure and use cases differ between different ledgers. By addressing this research gap, the primary goal of this work is to test the applicability of the selected research methods on XRP by adapting them to the specific properties of the networks' structure as well as the available off-chain data. The secondary goal is to provide insights into the implications of our results for the XRP ecosystem and potential implementation scenarios for our fraud detection classifiers.

The relevance to apply this research on the XRP Ledger arises from its increasing importance in financial markets. The lack of studies in this area creates ambiguity regarding the network's usage, activity types, and techniques to mitigate fraudulent behavior. This presents a challenge for regulators, compliance bodies, investors, financial institutions, and the broader XRP ecosystem in assessing the network's reliability and its participation in economically significant activities within markets. The segmentation of agents and detection of fraud in the network has the potential to implement effective ecosystem monitoring and demonstrate XRP's utility and diverse applications to stakeholders, including investors, financial institutions, and

researchers. The potential of the utilized methods also lies in monitoring regulatory compliance, especially in jurisdictions with oversight of typical high-risk activities. Ripple can leverage the findings of our work to optimize the network for its most valuable segments and manage funded partnerships through data-driven decisions. Eventually, the development of a model capable of detecting fraudulent accounts is expected to benefit regulatory and prosecutorial bodies by facilitating automated real-time detection of illicit agents, thereby strengthening compliance and enforcement mechanisms. The trust of users, particularly institutional stakeholders, is expected to depend on XRP being a secure payment environment where fraudulent activities are effectively identified and mitigated. Ultimately, we expect this work to contribute to greater transparency in a financially significant blockchain network, fostering trust, supporting informed decision-making, and enhancing confidence in its integrity.

2 The XRP Ledger

This section introduces the XRP ecosystem as the primary domain of this study. The XRP Ledger (XRPL) is a blockchain solution and distributed ledger technology (DLT) developed and first launched in 2012 by Ripple Inc (Ripple, n.d.-c). A blockchain is a decentralized public ledger that records transactions within a peer-to-peer network. It consists of linked data blocks, each referencing the previous one. The XRP network relies on trusted validator nodes to reach consensus on data validity, enabling secure and integer transactions without a central authority (Ripple 2024b; European Union Agency for Cybersecurity (ENISA), n.d.).

The XRPL leverages its native cryptocurrency, XRP. Unlike Bitcoin (BTC), which requires mining to create new coins, all 100 billion XRP tokens were created at its inception, ensuring a fixed supply from the beginning (Alahmad et al. 2023; Li and Whinston 2020; Ripple 2024b). As of November 2024, XRP's market capitalization exceeded $100 billion, with each token valued at around 1.80 USD (CoinMarketCap 2024). Ripple Inc., is the company and

issuer behind the network and token, holding about 48 billion XRP in escrow (2024), with a monthly release of up to 1 billion tokens to stabilize XRP supply and prices (Sergeenkov 2024). The company has partnered with over 200 financial institutions, including major and central banks, to integrate the XRPL network into their systems (Rella 2020; Ahmadova and Erek 2022).

Designed to enable seamless global value transfers, the XRPL primarily functions as a distributed payment system aimed at improving cross-border payment efficiency, offering an alternative to traditional solutions like SWIFT (Chase and MacBrough 2018; Sergeenkov 2024). While widely adopted, SWIFT transactions can be slow and costly due to intermediary fees and the need for pre-funded foreign accounts. In contrast, the XRPL supports rapid, cost-effective transactions across multiple currencies through features like trust lines and an on-chain decentralized exchange (Rella 2020; Peduzzi, James, and Xu 2021). By eliminating intermediary fees and the need for pre-funded foreign accounts, adopting the XRPL can reduce operational costs for banks by up to 60% (Sergeenkov 2024).

Beyond payments, the XRPL network supports key applications, including tokenization, central bank digital currencies (CBDCs), and stablecoins, as well as Non-Fungible tokens (NFTs) (Ripple, n.d.-b). In comparison, other ecosystems like BTC primarily serve as digital assets, while networks such as Ethereum (ETH) focus on decentralized applications powered by smart contracts (Alahmad et al. 2023).

Ultimately, the XRPL also differs from other blockchain technologies in its consensus mechanism and fee system. The XRPL employs the Ripple Protocol Consensus Algorithm (RPCA), where trusted validators from a Unique Node List (UNL) vote on transaction validity. This process confirms transactions within seconds, enabling fast and efficient processing. In comparison, BTC's Proof of Work (PoW) mechanism relies on miners solving complex problems, resulting in an average block time of ten minutes. ETH transitioned from the PoW in

2022, to a less energy-intensive approach called Proof of Stake (PoS) that uses coin ownership for validation (Mauri, Cimato, and Damiani 2018; Alahmad et al. 2023). Thereby, the XRPL outperforms other blockchains in terms of transaction speed and fees, processing around 1,500 transactions per second with costs of approx. 0.00001 XRP (depending on network load) per transaction, which represents only a fraction of a dollar cent. (Mauri, Cimato, and Damiani 2018; Alahmad et al. 2023; Sergeenkov 2024)

The XRPL also differs in its data and ledger structure. It uses an account-based system that directly tracks each account's balance, unlike BTC's Unspent Transaction Output (UTXO) model, where balances are spread across multiple addresses and transactions involve multiple inputs and outputs. This simplifies transaction validation, as XRPL accounts can spend part of their balance while retaining the rest, using a single address repeatedly for both sending and receiving funds. (Mauri, Cimato, and Damiani 2018; Alahmad et al. 2023; Akcora, Gel, and Kantarcioglu 2022)

Eventually, each XRP transaction includes detailed information such as the sender and receiver addresses, transaction amount, fee, transaction type (from over 40 types), and additional metadata such as source tags for user identification, optional memos for supplementary information, and flags or settings related to the transaction. (Ripple, n.d.-a; 2024a)

3 Research Context and Related Studies

This section reviews related works on characterizing economic activity types, segmenting accounts, and automating the detection of fraudulent users on DLTs. Section 3.1 provides an overview of agent segmentation, with Sections 3.2 focusing on the application of heuristics.

3.1 Segmenting Economic Agents on DLTs

The segmentation of agents on DLTs has been a prominent research topic in recent years. Thereby, economic agents can be defined as individuals, organizations, or entities engaged in economic activities, including decisions related to the production, distribution, and consumption of goods and services (United Nations, n.d.). Reviewing related works reveals that the definitions of economic agent segments in blockchain networks are highly adaptable, often influenced by the specific use cases of the ledger analyzed. It is reasonable to expect that the primary use cases of the XRPL align with its key applications, as outlined in Section 2. However, since no prior efforts have segmented users within its ecosystem, insights from other blockchains offer valuable guidance for identifying and segmenting XRPL agents. *Chainalysis* categorizes economic entities within blockchain ecosystems, such as services and organizations, based on their specific use cases of the technology (

Table *1*).

Category	**Description**
Merchant Services	Businesses that accept cryptocurrency as payment for goods or services, functioning similarly to traditional payment processors.
Hosted Wallets	Custodial services where the provider manages the security of users' public and private keys, making transactions easier but with less financial privacy.
Mining Pools	Groups that combine computing power to mine cryptocurrency, often considered low-risk as they primarily earn through mining rewards.
Exchanges	Platforms where users can buy, sell, or trade cryptocurrency. Includes both centralized (custodial) and decentralized (non-custodial) exchanges.
Nested Services	Services operating within larger exchanges, such as instant exchangers and OTC brokers, which can pose risks if compliance standards are lax.
Gambling Services	Platforms for betting and gaming. Risk level varies by jurisdiction; can be used for legitimate purposes or as a means for money laundering.
Cyber Infrastructure	Providers of web hosting, VPNs, and other internet-based services that accept cryptocurrency, sometimes offering anonymity features, leading to medium risk categorization.
Mixers	Services that enhance transaction privacy by mixing funds to obscure their origin and destination. Frequently scrutinized for potential use in money laundering.
Darknet Markets	Online platforms for trading illegal goods and services, accessible via anonymity networks like Tor. High-risk due to involvement in illicit activities.
Illicit Actor Organizations	Groups associated with activities such as ransomware, fraud, terrorism, or money laundering. Includes organizations indirectly linked to illegal activities.

Table 1: Chainalysis Blockchain Segments (Chainalysis, n.d.)

The categories include businesses accepting cryptocurrencies (merchant services), custodial wallets (hosted wallets), mining pools, centralized or decentralized trading platforms

(exchanges), nested services within exchanges, gambling sites, cyber-infrastructure providers, privacy-enhancing mixers, darknet markets, and entities linked to illegal activities. Thereby, *Chainalysis* does not provide any ledger-specific information, it mainly focuses on categorizing actors without diving into specific mechanisms unique to individual ledger ecosystems.

However, the diversity of blockchain ecosystems necessitates segmentation approaches tailored to the specific characteristics of each ledger. Blockchain use cases and ledger mechanisms differ significantly as mentioned in Section 2. The following subsections will present related works and their approaches in characterizing and identifying agents specific to unique ledger technologies.

3.2 Behavioral Analysis and Account Heuristics

The following chapter explores methodologies and related research relevant to classifying economic agents on DLTs through heuristics and behavioral analysis (overview in Table 16).

Wu et al. (2021) categorize methods for associating blockchain addresses with real-world entities – commonly called "de-anonymizing" – into three groups: transaction property-based, behavior-based, and off-chain information-based. De-anonymization, widely studied in BTC research, aims to link addresses to entities by analyzing transactional patterns. While our work does not focus on entity recognition, these methods are relevant as they overlap with approaches for segmenting and characterizing different groups of agents.

Heuristics play a central role in these methodologies as they simplify decision-making by focusing selectively on the most relevant information while disregarding less critical details (Gigerenzer and Gaissmaier 2011). These strategies allow researchers to analyze blockchain data more efficiently by simplifying complex decision-making processes. Heuristics are valuable for characterizing user groups by clustering addresses that exhibit similar transactional patterns and behaviors (Nick 2015). Beyond classification, heuristics are commonly used to

analyze the pseudo-anonymity of transactions and tracing the movement of funds within the blockchain (Sun Yin et al. 2019). By identifying recurring behavioral patterns, heuristics not only enhance our understanding of agent interactions but also enable the differentiation of economically significant activities from other forms of behavior (Harrigan and Fretter 2016).

Transaction Properties

This approach utilizes the structural characteristics of blockchain transactions to identify addresses likely controlled by the same entity. It facilitates the analysis of user behavior despite the pseudo-anonymity of these systems (Wu et al. 2021).

A well-known example for this is the multi-input heuristic in BTC, also known as common spending, which clusters addresses that act as inputs in the same transaction under the assumption that they belong to the same user (Meiklejohn et al. 2013; Ermilov, Panov, and Yanovich 2017). This heuristic leverages the structural requirements of BTC's transaction model to infer shared ownership and has been widely adopted in studies focusing on address clustering and network behavior, (Androulaki et al. 2013; Meiklejohn et al. 2013; Harrigan and Fretter 2016; Jourdan et al. 2018; Toyoda, Ohtsuki, and Mathiopoulos 2018; Nakamoto 2008).

Similarly, the one-time change (or "shadow" address) heuristic identifies change addresses created to receive leftover funds in a transaction. The heuristic assumes that these change addresses, often newly created, are managed by the same entity that controls the input addresses (Androulaki et al. 2013; Meiklejohn et al. 2013).

Beyond these foundational heuristics, additional techniques have been introduced to refine clustering and improve precision. Nick (2015) proposed heuristics targeting specific properties, such as consumer wallets defaulting on single-address outputs or optimizing change outputs for minimal fragmentation.

In the ETH ecosystem, the account-based model requires different heuristic strategies, reflecting the structural differences in how transactions and accounts are managed. Victor (2020) proposed grouping addresses based on deposit account reuse, airdrop participation, and token authorization patterns, reflecting the distinct transactional structures in ETH.

User Behavior

These methods analyze transaction patterns to uncover recurring behaviors and relationships between blockchain addresses. By examining features such as timing, frequency, and transaction amounts, these methods aim to group addresses on DLT networks (Wu et al. 2021).

Reid and Harrigan (2011) used transaction timing over extended periods to link entities and addresses. Their work showed that this approach can yield significant insights into agent activity and relationships, particularly when combined with external data sources. Similarly, Androulaki et al. (2013) employed features like transaction time, sender and receiver indices, and transaction amounts to cluster BTC users. Their findings revealed that these features could be used to identify nearly 40% of agents in their dataset, demonstrating the efficacy of behavior-based clustering for uncovering user roles and activities. Monaco (2015) examined long-term agent activity in BTC transactions and proposed different features to quantify behavioral dynamics. The study emphasized the role of timing and complex network attributes in capturing predictable patterns over time. These patterns were shown to reduce anonymity, demonstrating how sustained behavioral observations can reveal user activities and highlight economically significant activities.

Ultimately, these studies demonstrate how observing agent activity over time with behavior-based methods provides insights into the economic activities and interaction patterns of entities in DLT networks.

Off-chain Information

This method adds external data sources to complement blockchain analysis and identify entities beyond the limitations of on-chain data. Such additional data can include any external information and details about the address owner, like IP addresses, contact details or social media handles. Including this information enables linking blockchain addresses to real-world identities and uncover patterns of behavior not immediately visible on-chain (Wu et al. 2021).

Reid and Harrigan (2011) used forum posts to trace entities involved in theft cases, demonstrating how external sources can reveal blockchain address ownership. Similarly, Ermilov, Panov, and Yanovich (2017) combined off-chain data from social media, wallet explorer, etc. with on-chain analysis, creating a more robust clustering framework.

Beyond user-reported data, network-level information such as IP addresses has also been used to link blockchain addresses to specific users. Integrating off-chain data with blockchain transaction information significantly broadens the scope of entity identification, reducing anonymity and uncovering critical behavioral patterns (Wu et al. 2021).

Application on the XRPL

While transaction property-based heuristics are widely used and effective in UTXO ecosystems, such as BTC, many of these approaches cannot be directly applied to the XRPL due to its account-based structure. For example, heuristics that rely on tracking individual outputs or input-linking methods as well as one-time change are not applicable on the XRPL, where a transaction can only be made through a single account. This creates the need for an alternative heuristic method, better matched to the XRPL's ledger model.

Behavioral features offer a valuable perspective by highlighting patterns in transaction timing, frequency, and volume, which can reveal distinct activity types and roles of entities. By adding information from external sources, the on-chain activities can be linked to real-world entities and their activity patterns can be validated.

4 Methodology

This section provides a concise overview of the proposed methods and model setups, as well as the data utilized in this work, including preprocessing and feature engineering steps. Subsection 4.1 presents an overview of the main datasets as well as the features engineered. Subsection 4.2 provides detailed descriptions of the data subsets utilized, the labeled ground truth data, feature selection processes, and the proposed heuristics as well as their evaluation methods.

4.1 Data and Feature Engineering

The transactional dataset for this study comprises over 2 billion XRPL transactions spanning 36 million ledgers (ledger 50,000,000 to 86,000,000), collected in a decentralized manner over 1,618 days (from 2019-09-13 to 2024-02-16). The main source of this data is a publicly available dataset with transactions from a full history ripple node provided by XRPL Labs developer *Wietse Wind* (Wind 2024).

During preprocessing, only successful transactions were retained to capture actual value transfers and avoid double-counting, as failed transactions are often retried. The XRPL supports over 40 transaction types, many of which relate to account settings or niche applications (Ripple, n.d.-a). To ensure a focus on economically significant activities, the analysis included transaction types such as *Payment, OfferCreate, OfferCancel, and TrustSet*. Additionally, escrow-related transactions *(EscrowCreate, EscrowCancel, etc.*) and specific NFT operations (*NFTokenCreateOffer, NFTokenAcceptOffer, NFTokenMint, etc.*) were included to account for value-transfer and contract-related activities. The selected transaction types account for over 99% of all transactions in the dataset. Other types, such as *AccountDelete, AccountSet, and TicketCreate*, were excluded as they do not represent asset movement or economic activity.

Feature Engineering

A total of 55 features were created on account level and are presented in Table 19 in the Appendix. The foundation for engineering these features are related studies discussed in Section 3, along with features specifically adapted to the characteristics of the XRPL. The features involve general transaction metrics that provide insights into overall account activity, including transaction count, active periods, and value movements, such as total amounts sent, received, and net balance changes. Payment ratios, divided into outgoing and ingoing, reflect how frequently accounts send or receive payments. Offer-related ratios and trust set & escrow ratios capture account involvement in creating offers or managing trust lines and escrows. NFT ratios, based on NFT related transaction types, describe activities related to NFT creation and trading. Tag presence ratios show how often tags are used, indicating interactions with platform-based entities that use tags to link transactions to end users. Finally, interaction metrics, such as unique transaction partners and ratio of interaction with the 20 largest exchanges, assess the diversity and characteristics of an account's interactions.

Aggregating XRP transaction data at the address level posed several challenges. The high volume of transactions made computations demanding, while the diversity of transaction types added complexity in defining consistent features. This was addressed by focusing on the most relevant transaction categories, as outlined earlier. Also, significant variance in activity levels between highly active accounts and smaller accounts with minimal transactions posed challenges in drawing meaningful insights across all accounts equally. To address this, ratio-based features were introduced to normalize activity levels and ensure account comparability. Additionally, data incompleteness and missing values created complications, particularly due to missing information on transaction amounts for non-XRP token payments. To address this, token-specific features were developed to distinguish between XRP and non-XRP transactions, ensuring a clearer data representation.

Ultimately, segmenting agents and identifying fraud required slightly different feature configurations. Therefore, different feature combinations were selected for each method to achieve optimal results, as detailed in the following subsections of Section 4.

Ground Truth Data

To aid in developing agent heuristics and interpreting clusters, an additional dataset containing labeled addresses was created, referred to as the 'known domains dataset' in this work. The primary sources for this dataset included a comprehensive collection of account names along with their associated web domains and *X* (formerly *Twitter*) handles, obtained through the *XRPScan* API. Additionally, it incorporated a dataset of domains published by address holders on the XRPL in the *verified domain* transaction field, gathered from the *Bithomp* website. (https://xrplexplorer.com/en/domains, as of October 6, 2024). Most of the addresses provided by *XRPScan* were manually verified by sending small test transactions to confirm their authenticity. The domains were manually labeled according to the categories in Table 2. Domains with unclear distinctions, such as those combining exchange activities with wallet or NFT marketplace hosting, as well as domains that were untraceable online or apparently linked to fraudulent activities, were excluded from the data.

Category	Definition
Exchanges	Addresses associated with centralized or decentralized platforms used for trading digital assets.
Gambling/Gaming	Addresses linked to platforms that provide gambling or gaming services, including casinos, online betting, and gaming with potential financial stakes.
NFT	Addresses connected to Non-Fungible Token (NFT) creation, sales, or related platform activities.
Services	Addresses belonging to businesses that provide various services, such as financial services like payment providers, loan services, or wallet providers, as well as non-financial utilities like consulting services.
Bridging Services	Addresses associated with platforms or entities to transfer assets or tokens between different blockchains or networks.
Issuer	Addresses that issue tokens or act as a source for specific assets on the XRPL. (These addresses may belong to entities involved in exchange, gambling, NFT, or other services.)
Other	Addresses that do not fit into the above categories, such as donation accounts of non-profit organizations, communities, and sellers of physical goods.

Table 2: Categories of economic agents, used as labels in known domains dataset

A separate dataset containing fraudulent addresses, provided by *XRP Forensics*, was used for the supervised fraud detection models and will be referred to as the 'fraud dataset' in this work. The dataset includes 12,351 fraudulent XRP accounts, consisting of a combination of flagged and auto-traced cases. The accounts were involved in a wide range of fraudulent activities, primarily consisting of addresses related to *PlusToken* fraud cases (32.27%), spam (27.19%), theft (19.31%), and giveaway scams (17.32%) as well as a minority of other types, such as token scams and different *Ponzi* schemes. Definitions of these fraud types were presented in Section 3.2. The dataset does not include other forms of illegal activity, such as involvement in the trade of illegal goods, money laundering, or terrorist financing.

4.2 Heuristic Methods for Agent Analysis

As outlined in Section 3.1.1, existing research used various approaches for characterizing different agents on blockchain networks. To meet the specific requirements of the XRPL, we combined a behavior-based approach with off-chain information. This involved leveraging selected behavioral features derived from the feature set (Table 19 in Appendix) alongside labeled addresses, included in the known domains dataset. The primary objective of this analysis is to identify distinctive behavioral patterns for the agent categories described in the known domains dataset (Table 2). We used statistical and descriptive data analysis to uncover variations in behavioral and interactional patterns, which served as a base to define heuristics for each category. These heuristics were iteratively refined and tested to ensure they accurately represented observed behaviors and aligned with the external classifications provided by the ground truth data. Figure 1 shows an overview of the entire approach.

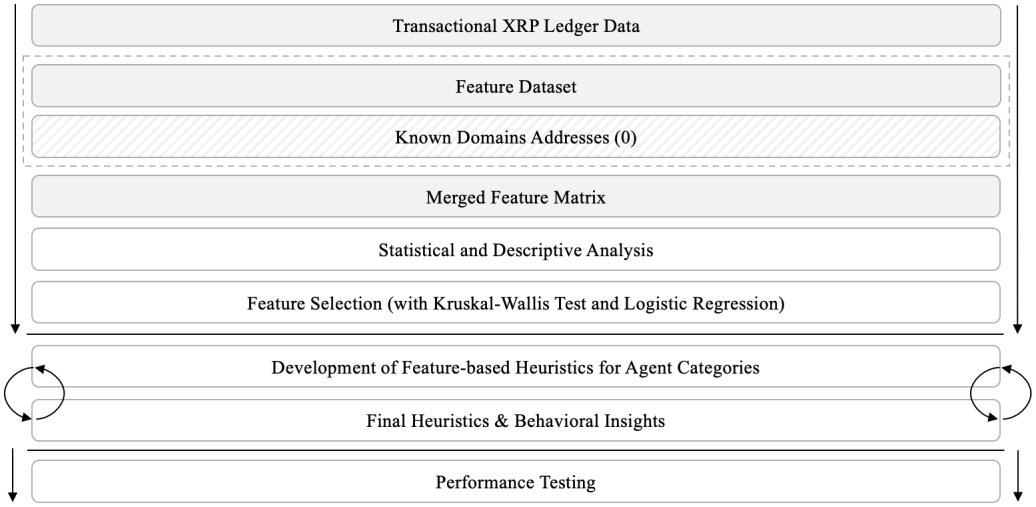

Figure 1: Methodological framework for heuristic-based agent analysis

Data

The feature matrix used for this analysis is a merged dataset, containing the features from the feature dataset for 2,130 labeled addresses from the known domains dataset. It contains following distribution of entries across agent categories: exchanges (1,465), gambling/gaming (64), NFTs (73), services (200), bridging services (32), issuers (73), and other (50).

To build an initial understanding, we examined transaction features such as frequency, transaction partners, and transaction amounts, which have proven effective for identifying behavioral patterns in cryptocurrency networks (Wu et al. 2021). Statistical summaries and visualizations were generated to explore data distributions, identify trends, and detect anomalies across categories. This was followed by an extensive exploratory data analysis (EDA) to examine feature distributions, correlations, and relationships, refining our understanding of behavioral patterns within the dataset.

Feature Selection

To identify features that effectively distinguish between the labeled categories, we applied the non-parametric Kruskal-Wallis test, suitable for non-normal and skewed distributions. This test

evaluates whether feature medians differ significantly across agent categories and measures their association strength (Ostertagová, Ostertag, and Kováč 2014).

Additionally, to identify the most influential features for each agent category, we encoded each category as a binary dependent variable and applied a systematic preprocessing workflow. Missing values were imputed with column means, and infinite values were replaced with NaN. Features with low variance were removed to eliminate non-informative predictors. To address multicollinearity, we calculated the Variance Inflation Factor (VIF) for all features and excluded those with VIF > 10. The remaining features were standardized to ensure numerical stability. Logistic regression with L2 regularization (Ridge) was then used to model the relationship between the features and each binary target variable. Regularization mitigated overfitting and addressed feature redundancy. The most influential features were determined based on the absolute values of the model coefficients. This process was repeated for all agent categories.

By combining the general features identified through the Kruskal-Wallis test with the most influential features from logistic regression, we expanded the feature pool to enhance its ability to differentiate and characterize each category effectively. This feature set served as the basis for developing heuristics tailored to identify distinctive patterns and behaviors of the various agents on the XRPL.

Proposed Heuristics Method

To characterize economic agents on the XRPL, we adapted existing heuristic methods (see Section 3.1.1.) to the specific properties of our dataset and the XRPL. The heuristics were defined to capture specific patterns in the data.

We started by analyzing the distribution of relevant features, which were previously identified in the feature selection. Statistical summaries, including mean, median, and standard

deviation, provided a baseline for setting initial thresholds. These thresholds were designed to capture the majority of addresses within a category while minimizing overlap with others.

The heuristics were then iteratively refined and tested on the feature matrix (features of known domains). Both individual features and combinations of multiple features were evaluated to identify the most effective rules. Performance was assessed using three key metrics:

- Recall measures the proportion of actual positive instances correctly identified, focusing on minimizing false negatives (FN), and is defined as $\frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$.
- False positive rate (FPR) measures the proportion of negative cases (accounts not belonging to a specific category) that are incorrectly classified as positive, defined as $\frac{False\ Positives\ (FP)}{False\ Positives\ (FP) + True\ Negatives\ (TN)}$
- F1 score is the harmonic mean of precision and recall, balancing both metrics.

The heuristics and their combinations were ranked by F1 scores, with a focus on maximizing recall and minimizing FPs to identify the most effective solutions. This ensured that the heuristics not only captured most target entities but also avoid misclassification.

To enhance performance, the top 10 heuristic (combinations) were refined in an iterative process. Threshold values were manually fine-tuned based on observed data distributions, standard deviations, and ranges to reduce overlap with other categories and improve coverage of the target category. Each adjustment was re-tested, and thresholds yielding improved performance replaced the previous values.

For consistently underperforming heuristics, additional features were tested. These new heuristics underwent the same threshold optimization and testing process, with performance compared across iterations. Refinement continued until further improvements plateaued, indicating that the heuristics had reached their optimal configuration.

5. Results and Discussion

The following subsection provides an initial overview of central behavioral patterns observed across the different agent categories. This is followed by a presentation of the feature selection results, which inform the heuristics development. In the last part of the section, the best-performing heuristics for each category are described in detail.

Behavioral Patterns: Issuers

Issuers display the lowest transaction frequencies across all categories, with minimal variability and a small range, suggesting more infrequent activity (Figure 2). Their *average_sent_amount_xrp* is generally lower (median: 24.66 XRP, mean: 4,088 XRP), reflecting smaller transaction sizes. In terms of interactions, issuers show the highest unique_transaction_partner_ratio (mean: 0.69, median: 0.83) with low variability. Issuers also exhibit the lowest *payment_with_XRP_ratio* on average (mean: 0.15, median: 0.03). Their *payment_as_destination_ratio* is higher compared to other categories, suggesting that the issuers in our data are more frequently on the receiving end of transactions.

Behavioral Patterns: NFT

NFT accounts exhibit behavioral patterns characterized by generally lower activity and engagement. They display lower transaction frequencies, with smaller average sent and received amounts compared to other categories. Additionally, NFT accounts show a low *payment_with_XRP_ratio* and low proportions of small-value payments, both as senders and receivers. The biggest difference can be seen in their concentration in NFT-specific features, such as the *nftoken_mint_ratio*, where they display a higher level of activity compared to other categories.

Behavioral Patterns: Exchanges

The exchanges in our dataset demonstrate a wide range of transaction frequencies, with a notable skew toward lower frequencies but also extreme outliers. Among all categories, exchanges exhibit the highest maximum of *transaction_frequency* and significant variability. They also record the highest sent and received amounts, with a mean sent amount of 8.62 million XRP and an extreme range (0.002–511.8 million XRP), surpassing all other categories both in average values and overall range. Additionally, exchanges maintain the highest payment-with-XRP ratio across all categories, with low variability.

Behavioral Patterns: Gambling/Gaming

Gambling and gaming accounts exhibit the highest average transaction frequencies and total transaction counts across all categories. Transaction amounts are generally small, but variability arises due to occasional larger transactions, with sent amounts reaching up to 173,231 XRP and received up to 8,152 XRP. They also demonstrate the lowest *unique_transaction_partner_ratio* (mean: 0.09, median: 0.02) among all categories. Similar to exchanges, gambling accounts exhibit a higher *payment_with_XRP_ratio.* Furthermore, the accounts in our dataset are characterized through a high *payment_as_destination_ratio* and *destination_tag_present_ratio.*

Behavioral Patterns: Services

The service accounts in our dataset display generally a higher *transaction_frequency*, with a wide range and significant variability, including minimal activity and extreme outliers. Sent and received amounts are among the largest across all categories, with average sent values of 52,403 XRP and received averages reaching 375,581 XRP. In terms of interactions, services have a moderate *unique_transaction_partner_ratio* (mean: 0.30), but high variability indicates substantial differences in engagement patterns among different service accounts.

Behavioral Patterns: Bridging services

Bridging service accounts display relatively consistent moderate *transaction_frequency* compared to other categories. Sent and received transaction amounts are substantial, with high medians and moderate variability. The average sent amount is over 15,500 XRP, while the received average reaches 53,733 XRP (Figure 2). The bridging services in our dataset also show moderate diversity in counterparty interactions, with a *unique_transaction_partner_ratio*.

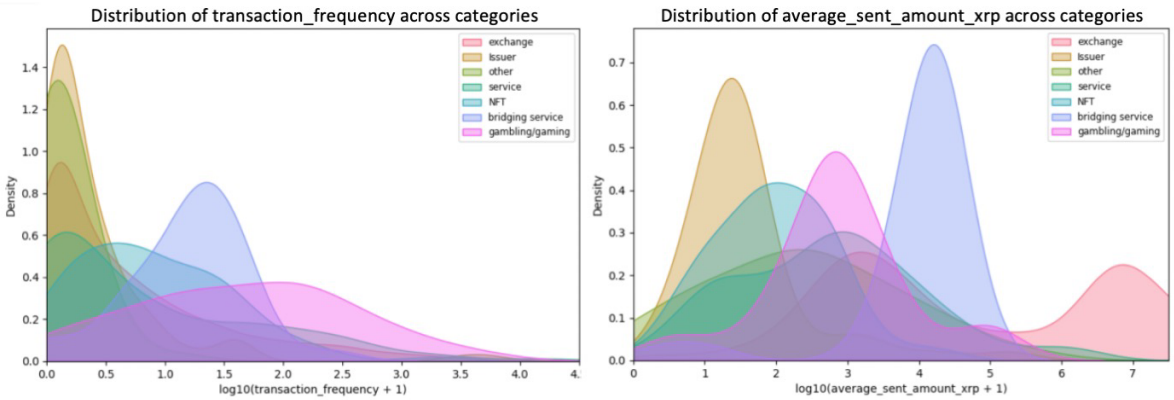

Figure 2: Distribution of transaction frequency and average sent amount across categories

The other category has been omitted due to the lack of clear behavioral patterns and the inability to assign them to specific entities. More visualizations of the data distribution of individual features are presented in Figure 3 in the Appendix.

Feature Selection for Heuristics

Based on the results of the Kruskal-Wallis test, we selected seven features that exhibit the most significant differences across categories:

Feature	**H-statistics**
nftoken_mint_ratio	H = 1285.59
payment_without_XRP_ratio	H = 788.17
payment_over_1000_xrp_receiver_ratio	H = 630.60
nftoken_create_offer_ratio	H = 612.53
payment_over_10000_xrp_receiver_ratio	H = 560.89
nftoken_cancel_offer_ratio	H = 538.62
nftoken_accept_offer_ratio	H = 480.57

Table 3: Top seven features based on Kruskal-Wallis Test

The test generates an H-statistic, which reflects the degree of variability among the medians of the groups being compared; a higher H-statistic indicates stronger differences between the groups (Ostertagová, Ostertag, and Kováč 2014). All seven selected features demonstrated p-values below the standard significance threshold of 0.05, confirming their statistic relevance (complete test results in Table 20).

Additionally, the logistic regression results revealed several features with influence on the likelihood of accounts belonging to specific agent categories (Table 21 and Table 22). Both positively and negatively associated features were identified, which indicate features that increase or decrease the likelihood of classification into a category. Features with odds ratios deviating significantly from 1 were considered most relevant and therefore added to the feature pool.

Final Heuristics: Issuers

Table 4 presents the most relevant heuristics (based on the overall test results) identified for the issuer category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
H_i1	*payment_as_destination_ratio* > 0.8	Over 80% of an account's transactions where it acts as receiver.
H_i2	*payment_with_XRP_ratio* < 0.2	Less than 20% of an account's payment transactions specify an XRP amount.
H_i3	*payment_without_XRP_ratio > 0.8*	Over 80% of an account's payment transactions do not specify an XRP amount.
H_i4	*unique_transaction_partner_ratio > 0.5*	Over 60% of an account's transactions involve unique partners.

Table 4: Most relevant heuristics for issuers

After testing all heuristics individually and different combinations, we found that the combination of *payment_as_destination_ratio* & *payment_with_XRP_ratio* was the best performing combination (Table 5). These results indicate that the issuers in our dataset can be characterized by their tendency to act as receiver of transactions and to conduct them without

XRP. Equally good performing was the second combination of *payment_as_destination_ratio* with *unique_transaction_partner_ratio* and *payment_with_XRP_ratio*. Although, adding the condition for a moderate diversity of transaction partners did not increase overall performance. The most effective heuristics for issuer accounts rely on a combination of low XRP usage, a high ratio of transactions where the account acts as the destination, and moderate transaction partner diversity.

Combination	Recall	False Positive Rate	F1 score
H_i1 and H_i2	0.60	0.003	0.72
H_i1 and H_i2 and H_i3	0.60	0.003	0.72
H_i1 and H_i3	0.55	0.003	0.68

Table 5: Best performing combination of heuristics for issuers

Final Heuristics: NFTs

Table 6 presents the most relevant heuristics (based on the overall test results) identified for the NFT category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
H_n1	*nftoken_mint_ratio* > 0.4	Over 40% of an account's transactions are NFT minting operations.
H_n2	*payment_without_XRP_ratio* < 0.2	Less than 20% of an account's payment transactions do not specify an XRP amount.
H_n3	*payment_over_1000_xrp_receiver* < *0.01*	Less than 1% of an account's received payments are over 1,000 XRP.

Table 6: Most relevant heuristics for NFTs

Table 7 shows that the single heuristic *nftoken_mint_ratio* achieved the best overall performance. This indicates that the proportion of minting transactions is a strong standalone feature for characterizing NFT accounts. The best performing combination consisted of *nftoken_mint_ratio* & *payment_without_XRP_ratio*, achieving the same performance metrics. Similarly, the combination of *nftoken_mint_ratio* with *payment_over_1000_xrp_receiver* did not affect the obtained performance. These capture behavioral patterns characterized by

frequent minting operations, a low proportion of non-XRP payments, or a minimal share of received payments exceeding 1,000 XRP.

Combination	Recall	False Positive Rate	F1 score
H_n1	0.63	0.002	0.75
H_n1 and H_n2	0.63	0.002	0.75
H_n1 and H_n3	0.63	0.002	0.75

Table 7: Best performing combination of heuristics for NFT

Final Heuristics: Exchanges

Table 8 presents the most relevant heuristics (based on the overall test results) identified for the exchange category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
H_e1	*payment_with_XRP_ratio* > 0.9	Over 90% of an account's payment transactions specify an XRP amount.
H_e2	*payment_without_XRP_ratio* < 0.1	Less than 10% of an account's payment transactions do not specify an XRP amount.
H_e3	*trust_set_ratio* < 0.01	Less than 1% of an account's transactions involve trustset operations.
H_e4	*nftoken_mint_ratio* < 0.01	Less than 1% of an account's transactions involve minting NFTs.
H_e5	*nftoken_create_offer_ratio* < 0.01	Less than 1% of an account's transactions involve creating NFT offers.
H_e6	*offer_create_with_XRP_ratio* < 0.001	Less than 0.1% of an account's transactions involve creating offers with specified XRP amounts.
H_e7	*offer_cancel_ratio* < 0.001	Less than 0.1% of an account's transactions involve canceling offers.
H_e8	*nftoken_accept_offer_ratio* < 0.01	Less than 1% of an account's transactions involve accepting NFT offers.
H_e9	*nftoken_cancel_offer_ratio* < 0.01	Less than 1% of an account's transactions involve canceling NFT offers.
H_e10	*escrow_ratio* < 0.01	Less than 1% of an account's transactions involve escrow operations.

Table 8: Most relevant heuristics for exchanges

Table 9 shows that the combination of all 10 exchange heuristics performed best. Equally well performing was the *payment_with_XRP_ratio* heuristic on its own. The results show that the *payment_with_XRP_ratio* heuristic is the most reliable indicator for characterizing exchange accounts in our dataset, suggesting their operational focus on XRP transactions when utilizing

the XRPL network. The high recall of these heuristics makes them effective at capturing the majority of exchange accounts.

Combination	Recall	False Positive Rate	F1 score
Full Combination (H_e1, H_e2, ... , H_e10)	0.98	0.38	0.93
H_e1	0.98	0.38	0.93
H_e1 and H_e2	0.98	0.38	0.93

Table 9: Best performing combination of heuristics for exchanges

Final Heuristics: Gambling & Gaming

Table 10 presents the most relevant heuristics (based on the overall test results) identified for the gambling/gaming category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
H_g1	*unique_transaction_partner_ratio* < 0.2	Less than 20% of an account's transactions involve unique partners.
H_g2	*total_transactions > 10,000*	More than 10,000 transactions conducted by the account.
H_g3	*payment_under_500_xrp_receiver_ratio > 0.4*	Over 40% of an account's payment transactions received are under 500 XRP.
H_g4	*payment_under_10_xrp_sender_ratio* < 0.2	Less than 20% of an account's payment transactions involve sending amounts under 10 XRP.

Table 10: Most relevant heuristics for gambling & gaming

The best-performing heuristic was a combination of *unique_transaction_partner_ratio* & *total_transactions,* shown in Table 11. The second-best heuristic combination was *unique_transaction_partner_ratio* & *payment_under_500_xrp_receiver_ratio.* The results suggest that gambling accounts in our data are characterized by a combination of frequent transactions, a narrow pool of transaction partners, and a high prevalence of small-value payments. These patterns suggest the assumption that gambling platforms typically involve repetitive, low-value transactions between users and the gambling service. While the heuristics effectively capture these behaviors, the moderate F1 scores indicate room for improvement.

Combination	Recall	False Positive Rate	F1 score
H_g1 and H_g2	0.58	0.05	0.38

H_g1 and H_g3	0.77	0.08	0.37
H_g2 and H_g4	0.63	0.06	0.37

Table 11: Best performing combination of heuristics for gambling/gaming

Final Heuristics: Services

Table 12 presents the most relevant heuristics (based on the overall test results) identified for the service category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
H_s1	*payment_without_XRP_ratio* > 0.36	Over 36% of an account's payment transactions do not specify an XRP amount.
H_s2	*payment_over_1000_xrp_receiver_ratio* < 0.1	Less than 10% of an account's payment transactions involve receiving over 1,000 XRP.
H_s3	*payment_over_10000_xrp_receiver_ratio* < 0.02	Less than 2% of an account's payment transactions involve receiving over 10,000 XRP.

Table 12: Most relevant heuristics for services

The best-performing heuristic combination was *payment_without_XRP_ratio* & *payment_over_1000_xrp_receiver_ratio*, shown in Table 13. Combining *payment_without_XRP_ratio* with *payment_over_10000_xrp_receiver_ratio*, achieved the same performance metrics as the first combination. The single heuristic *payment_without_XRP_ratio* ranked third. These results reflect the transactional behavior of service accounts in our data, which often engage in smaller transactions while relying less on XRP. The relatively low FPR suggests that these heuristics are effective in narrowing down potential service accounts. However, the moderate identification F1 scores indicate the potential for refining these heuristics further to improve their accuracy.

Combination	Recall	False Positive Rate	F1 score
H_s1 and H_s2	0.45	0.06	0.46
H_s1 and H_s3	0.45	0.06	0.46
H_s1	0.45	0.06	0.45

Table 13: Best performing combination of heuristics for services

Final Heuristics: Bridging Services

Table 14 presents the most relevant heuristics (based on the overall test results) identified for the bridging service category out of all heuristics that were defined and evaluated during the analysis.

Name	Heuristic	Description
$H_{bs}1$	*average_sent_amount_xrp* between 1,000 and 316,227 XRP	Average amount sent per transaction is between 1,000 and 316,227 XRP.
$H_{bs}2$	*average_received_amount_xrp* between 1,000 and 316,227 XRP	Average amount received per transaction is between 1,000 and 316,227 XRP.
$H_{bs}3$	*absolute_active_period* between 3 and 100 days	Account activity spans between 3 and 100 days.
$H_{bs}4$	*total_transactions* between 31.6 and 31,622.8	Total transaction count is between 31.6 and 31,622.8.
$H_{bs}5$	*stddev_sent_amount_xrp* between 10,000 and 3,162,277.6 XRP	Standard deviation of sent transaction amounts is between 10,000 and 3,162,277.6 XRP.
$H_{bs}6$	*stddev_received_amount_xrp* between 10,000 and 10,000,000 XRP	Standard deviation of received transaction amounts is between 10,000 and 10,000,000 XRP.
$H_{bs}7$	*unique_destination_partners > 10*	Account interacts with more than 10 unique destination partners.

Table 14: Most relevant heuristics for bridging services

The best-performing heuristic combinations were *average_sent_amount_xrp* & *absolute_active_period* and *average_received_amount_xrp* & *absolute_active_period.* The combination of all heuristic also performed well but achieved a significantly lower recall than the other two combinations (Table 15). These results highlight that bridging service accounts in our data are characterized by significant average transaction amounts, moderate activity periods, and diverse interactions. The precision of the heuristics suggests that bridging services operate in a distinct transactional space, balancing high-value movements with a limited but consistent range of activity durations. The moderately high F1 scores indicate that these heuristics effectively capture bridging service behaviors, although there is room for refinement to further minimize FPs while maintaining strong recall.

Combination	Recall	False Positive Rate	F1 score
$H_{bs}1$ and $H_{bs}3$	0.72	0.006	0.69
$H_{bs}2$ and $H_{bs}3$	0.72	0.006	0.69
Full Combination ($H_{bs}1$, $H_{bs}2$, ... , $H_{bs}10$)	0.45	0.06	0.45

Table 15: Best performing combination of heuristics for bridging services

Practical Implications

The application of account-based heuristics proved to be an effective method for characterizing different economic agent categories, with varying degrees of success depending on the complexity of each category.

For NFT accounts, a single heuristic *(nftoken_mint_ratio)* achieved the best performance, showing how distinct behaviors can be easily captured using a focused heuristic. Similarly, the analysis achieved high performance with just two heuristics for issuer accounts and bridging services.

For exchange accounts, the highest recall and F1 score were achieved simultaneously with either 10 heuristics or just one. However, the relatively high FPR shows the difficulty in isolating exchange accounts due to overlapping behaviors with other account types. Gambling/gaming and service accounts were more difficult to characterize.

The ability of heuristics to efficiently capture distinct transaction behaviors can be applied in real-world applications. While some categories can already be effectively characterized, the challenges observed with other categories show the importance of refining these methods for further analysis. With a growing volume of off-chain data, the accuracy of heuristics is expected to improve, enhancing their generalization capabilities. Since the XRPL remains relatively unexplored, heuristics offer researchers a valuable initial step toward understanding economic activities on the network, aligning with the typical process of early ledger exploration (Section 3.1.1). In the future, heuristics could support the development of larger labeled datasets, which can be leveraged for supervised machine learning techniques, as demonstrated by researchers on other DLTs (e.g., Lin et al., 2019). While supervised machine learning was not implemented in this study due to the limited availability of XRPL-specific data, it generally represents a powerful tool for precisely segmenting economic agents.

Advancing these methods on the XRPL would enable more targeted and detailed insights into the behaviors of diverse agent categories.

6 Limitations

Our study is subject to several limitations that impact the scope and depth of the analysis.

The feature dataset was restricted to a specific time frame due to computational constraints, preventing the examination of transaction patterns beyond this period. The random sampling of data used for different approaches introduces variability that could impact the consistency of results. Crucial metadata, such as details of non-XRP token transactions, is absent due to the limitations of the public data source used. The study is also limited by its feature set. Time-based variables could not be fully computed due to computational resource constraints, and account-specific metrics, such as total address balances, were unavailable. Eventually, the inclusion of graph-based features could have enhanced the analysis by providing deeper insights into relationships and interactions between accounts.

Additionally, self-labeling the known domains data might have introduced inaccuracies or inconsistencies, potentially biasing the results. Furthermore, entities often possess multiple addresses, which they presumably use for different purposes, leading to varied transaction patterns across these addresses. This makes characterizing and verifying segments challenging.

Ultimately, the dynamic nature of blockchain usage presents a hurdle. Methodologies effective today may become less applicable over time as user behaviors and network structures evolve (Victor 2020).

Further limitations specific to our distinct approaches include:

In the case of heuristics, future work could explore a broader feature set to enhance differentiation. While maintaining simplicity, this could help to reduce overlaps between entity types. The utilized know domains dataset of 2,130 labeled addresses is limited in size and

imbalanced, with smaller categories like bridging services having too few addresses to ensure statistically reliable results. All tests and manually set thresholds were limited to this data, which affects its reliability when modeling unknow data with different distributions and dynamics. In consequence, working with larger and more balanced datasets could improve reliability and applicability.

Eventually, the feature selection approach and the use of statistical summaries (e.g., means, medians) may oversimplify complex interactions between features. Selecting the best features and setting appropriate thresholds is fundamental to the effectiveness of the heuristics. Incorporating more advanced methods, such as those accounting for non-linear patterns could uncover more nuanced insights and improve accuracy.

7 Conclusion

By analyzing XRPL transaction data on account-level, this study has shown that the segmenting XRPL agents is partially doable and can inform on the characteristics of the economic activity inherent in the XRP network. Furthermore, the study demonstrated the capability of supervised machine learning in effectively detecting fraudulent XRP accounts.

The need for greater transparency on the XRPL arises from its role in facilitating large-scale cross-border transactions, token issuance, decentralized exchange activities, NFT operations, all conducted in a pseudonymous manner. This lack of transparency presents challenges for regulators, compliance bodies, investors, financial institutions, and the broader XRP ecosystem in assessing the network's reliability and its participation in economically significant activities within markets.

Despite the scarcity of ground truth off-chain data and the lack of research on XRPL network activity, our study successfully uncovered novel insights and introduced a framework for distinguishing economic agent categories on the XRPL.

Our account-based heuristics establish a foundational framework for characterizing different economic agent categories on the ledger, providing a hypothesis-driven methodology that highlights distinct patterns within the ground truth data. This makes it particularly valuable for simplifying complex data structures and gaining initial insights into the XRP network.

Eventually, certain limitations of our study must be acknowledged. Constraints in data availability, including limited metadata and imbalances in the ground through datasets, affected the generalizability of results. Computational restrictions limited the inclusion of more complex features and larger data subsets, which could have enhanced insights into account behaviors and interactions. Additionally, our self-labeled data might have introduced some bias as well as the assumptions made during data sampling processes may influence the reliability of the findings. Future work should address these limitations by leveraging more extensive datasets, incorporating advanced feature engineering, and adapting the methodologies to the ever evolving blockchain dynamics.

Ultimately, the insights and methodologies developed in this study extend beyond the XRPL, offering a framework that can be adapted to other blockchain ecosystems to address similar challenges in transparency, fraud detection, and user segmentation. By bridging on-chain analysis with off-chain validation, reliable insights can be generated to enhance trust and accountability in decentralized systems. To build on our findings, further research should prioritize collaborations between blockchain developers, regulatory bodies, and academic institutions to refine data collection processes and standardize methodologies for addressing fraud and economic activity monitoring. Strengthening transparency and mitigating risks in blockchain ecosystems will be key to enforce broader adoption and acceptance of decentralized technologies in financial markets, and beyond.

List of References

Ahmadova, Sevinj, and Mustafa Salim Erek. 2022. 'A Review on Ripple, a Financial Intermediary Coin'. *Journal of Academic Projection* 7 (2): 117–30.

Akcora, Cuneyt Gurcan, Yulia R. Gel, and Murat Kantarcioglu. 2022. 'Blockchain Networks: Data Structures of Bitcoin, Monero, Zcash, Ethereum, Ripple, and Iota'. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery* 12 (1): e1436. https://doi.org/10.1002/widm.1436.

Alahmad, Mohammed, Adel Alfouderi, Ahmad Alonaizi, and Meshal Aldhamen. 2023. 'Comparison Study of the Top 5 Leading Cryptocurrencies Based on General Consensus Protocol: Bitcoin, Ethereum, Tether, XRP and Bitcoin Cash'. *WSEAS TRANSACTIONS ON COMPUTER RESEARCH* 11 (April):23–32. https://doi.org/10.37394/232018.2023.11.3.

Androulaki, E, G. O. Karame, M. Roeschlin, T. Scherer, and S. Capkun. 2013. 'Evaluating User Privacy in Bitcoin'. In , Revised Selected Papers 17:34–51. Okinawa, Japan: Springer Berlin Heidelberg.

Birrane, Kieran Daniel. n.d. 'An Exploration of Blockchain: An Unsupervised Analysis of the Ethereum Network'.

Breiman, Leo. 2001. 'Random Forests'. *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Caliński, T., and J Harabasz. 1974. 'A Dendrite Method for Cluster Analysis'. *Communications in Statistics* 3 (1): 1–27. https://doi.org/10.1080/03610927408827101.

Chainalysis. 2019. 'PlusToken Scammers Didn't Just Steal $2+ Billion Worth of Cryptocurrency. They May Also Be Driving Down the Price of Bitcoin.' *Chainalysis* (blog). 16 December 2019. https://www.chainalysis.com/blog/plustoken-scam-bitcoin-price/.

———. 2024. 'The 2024 Crypto Crime Report'. https://www.chainalysis.com/wp-content/uploads/2024/06/the-2024-crypto-crime-report-release.pdf.

———. n.d. 'Key Players In Crypto Report'. Chainalysis.

Chase, Brad, and Ethan MacBrough. 2018. 'Analysis of the XRP Ledger Consensus Protocol'. arXiv. https://doi.org/10.48550/arXiv.1802.07242.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. 'SMOTE: Synthetic Minority Over-Sampling Technique'. *Journal of Artificial Intelligence Research* 16 (June):321–57. https://doi.org/10.1613/jair.953.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco California USA: ACM. https://doi.org/10.1145/2939672.2939785.

CoinMarketCap. 2024. 'Cryptocurrency Prices, Charts And Market Capitalizations'.

CoinMarketCap. 2024. https://coinmarketcap.com/.

Davies, David, and Don Bouldin. 1979. ‘A Cluster Separation Measure’. *Pattern Analysis and Machine Intelligence, IEEE Transactions On* PAMI-1 (May):224–27. https://doi.org/10.1109/TPAMI.1979.4766909.

Ermilov, Dmitry, Maxim Panov, and Yury Yanovich. 2017. ‘Automatic Bitcoin Address Clustering’. In , 461–66. https://doi.org/10.1109/ICMLA.2017.0-118.

European Union Agency for Cybersecurity (ENISA). n.d. ‘Blockchain’. Page. ENISA. Accessed 28 November 2024. https://www.enisa.europa.eu/topics/incident-response/glossary/blockchain.

Farrugia, Steven, Joshua Ellul, and George Azzopardi. 2020. ‘Detection of Illicit Accounts over the Ethereum Blockchain’. *Expert Systems with Applications* 150 (July):113318. https://doi.org/10.1016/j.eswa.2020.113318.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. ‘Heuristic Decision Making’. *Annual Review of Psychology* 62 (Volume 62, 2011): 451–82. https://doi.org/10.1146/annurev-psych-120709-145346.

Harlev, Mikkel Alexander, Haohua Sun Yin, Klaus Christian Langenheldt, Raghava Rao Mukkamala, and Ravi Vatrapu. 2018. ‘Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning’. In . https://core.ac.uk/download/pdf/143481278.pdf.

Harrigan, Martin, and Christoph Fretter. 2016. *The Unreasonable Effectiveness of Address Clustering*. https://doi.org/10.48550/arXiv.1605.06369.

Hellwig, Daniel, Goran Karlic, and Arnd Huchzermeier. 2020. *Build Your Own Blockchain: A Practical Guide to Distributed Ledger Technology*. Management for Professionals. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-40142-9.

Jourdan, Marc, Sebastien Blandin, Laura Wynter, and Pralhad Deshpande. 2018. *Characterizing Entities in the Bitcoin Blockchain*. https://doi.org/10.48550/arXiv.1810.11956.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. ‘LightGBM: A Highly Efficient Gradient Boosting Decision Tree’. In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Kılıç, Baran, Alper Sen, and Can Özturan. 2022. ‘Fraud Detection in Blockchains Using Machine Learning’. In *2022 Fourth International Conference on Blockchain Computing and Applications (BCCA)*, 214–18. https://doi.org/10.1109/BCCA55292.2022.9922045.

Kraken. n.d. ‘Beware of Crypto Giveaway Scams | Kraken’. Accessed 27 November 2024. https://support.kraken.com/hc/en-us/articles/360057159411-Beware-of-crypto-giveaway-scams.

Li, Xiaofan, and Andrew Whinston. 2020. ‘Analyzing Cryptocurrencies’. *Information Systems Frontiers* 22 (February). https://doi.org/10.1007/s10796-019-09966-2.

Lin, Yu-Jing, Po-Wei Wu, Cheng-Han Hsu, I-Ping Tu, and Shih-wei Liao. 2019. ‘An Evaluation of Bitcoin Address Classification Based on Transaction History Summarization’. In *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 302–10. https://doi.org/10.1109/BLOC.2019.8751410.

Lorenz, Joana, Maria Inês Silva, David Aparício, João Tiago Ascensão, and Pedro Bizarro. 2021. ‘Machine Learning Methods to Detect Money Laundering in the Bitcoin Blockchain in the Presence of Label Scarcity’. arXiv. http://arxiv.org/abs/2005.14635.

Lundberg, Scott, and Su-In Lee. 2017. ‘A Unified Approach to Interpreting Model Predictions’. arXiv. http://arxiv.org/abs/1705.07874.

Mauri, Lara, Stelvio Cimato, and Ernesto Damiani. 2018. ‘A Comparative Analysis of Current Cryptocurrencies’: In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 127–38. Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0006648801270138.

Meiklejohn, Sarah, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. 2013. ‘A Fistful of Bitcoins: Characterizing Payments among Men with No Names’. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, 127–40. Barcelona Spain: ACM. https://doi.org/10.1145/2504730.2504747.

Molnar, Christoph. 2024. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/feature-importance.html.

Monaco, John V. 2015. ‘Identifying Bitcoin Users by Transaction Behavior’. In , edited by Ioannis A. Kakadiaris, Ajay Kumar, and Walter J. Scheirer, 945704. Baltimore, Maryland, United States. https://doi.org/10.1117/12.2177039.

Morgia, Massimo La, Alessandro Mei, Francesco Sassi, and Julinda Stefa. 2024. ‘Pump and Dumps in the Bitcoin Era: Real Time Detection of Cryptocurrency Market Manipulations’. arXiv. http://arxiv.org/abs/2005.06610.

Nakamoto, Satoshi. 2008. ‘Bitcoin: A Peer-to-Peer Electronic Cash System’.

Nerurkar, Pranav, Sunil Bhirud, Dhiren Patel, Romaric Ludinard, Yann Busnel, and Saru Kumari. 2021. ‘Supervised Learning Model for Identifying Illegal Activities in Bitcoin’. *Applied Intelligence* 51 (June):1–20. https://doi.org/10.1007/s10489-020-02048-w.

Nick, Jonas David. 2015. ‘Data-Driven De-Anonymization in Bitcoin’. Application/pdf,Online-Ressource. https://doi.org/10.3929/ETHZ-A-010541254.

Ostapowicz, Michał, and Kamil Żbikowski. 2019. ‘Detecting Fraudulent Accounts on Blockchain: A Supervised Approach’. In , edited by Reynold Cheng, Nikos Mamoulis,

Yizhou Sun, and Xin Huang, 11881:18–31. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-34223-4_2.

Ostertagová, Eva, Oskar Ostertag, and Jozef Kováč. 2014. 'Methodology and Application of the Kruskal-Wallis Test'. *Applied Mechanics and Materials* 611:115–20. https://doi.org/10.4028/www.scientific.net/AMM.611.115.

Payette, James, Samuel Schwager, and Joseph Murphy. 2017. 'CHARACTERIZING THE ETHEREUM ADDRESS SPACE'.

Peduzzi, Gaspard, Jason James, and Jiahua Xu. 2021. 'Jack the Rippler: Arbitrage on the Decentralized Exchange of the XRP Ledger'. In *2021 3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, 1–2. https://doi.org/10.1109/BRAINS52497.2021.9569833.

Ramezan, Gholamreza, and Cyril Leung. 2020. 'Analysis of Proof-of-Work-Based Blockchains Under an Adaptive Double-Spend Attack'. *IEEE Transactions on Industrial Informatics* 16 (11): 7035–45. https://doi.org/10.1109/TII.2020.2977689.

Reid, Fergal, and Martin Harrigan. 2011. 'An Analysis of Anonymity in the Bitcoin System'. *Security and Privacy in Social Networks* 3 (July). https://doi.org/10.1109/PASSAT/SocialCom.2011.79.

Rella, Ludovico. 2020. 'Steps towards an Ecology of Money Infrastructures: Materiality and Cultures of Ripple'. *Journal of Cultural Economy* 13 (2): 236–49. https://doi.org/10.1080/17530350.2020.1711532.

Ripple. 2024a. 'Transaction Metadata'. 9 October 2024. https://xrpl.org/docs/references/protocol/transactions/common-fields.

———. 2024b. 'What Is the XRP Ledger?' https://xrpl.org/docs/introduction/what-is-the-xrp-ledger.

———. n.d.-a. 'Transaction Types'. XRPL. Accessed 6 November 2024. https://xrpl.org/docs/references/protocol/transactions/types.

———. n.d.-b. 'XRP Ledger - Use Cases & Featured Projects'. Accessed 28 November 2024. https://xrpl.org/about/uses.

———. n.d.-c. 'XRP Ledger History'. Accessed 28 November 2024. https://xrpl.org/about/history.

———. n.d.-d. 'XRP Ledger Home | XRPL.Org'. Accessed 30 November 2024. https://xrpl.org/docs/concepts/tokens/fungible-tokens/authorized-trust-lines.

Rousseeuw, Peter J. 1987. 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis'. *Journal of Computational and Applied Mathematics* 20 (November):53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Scikit-Learn. 2024. 'Scikit Learn Documentation: Decision Trees'. 2024. https://scikit-

learn/stable/modules/tree.html.

———. n.d.-a. ‘Ensembles: Gradient Boosting, Random Forests, Bagging, Voting, Stacking’. Scikit-Learn. Accessed 7 November 2024. https://scikit-learn/stable/modules/ensemble.html.

———. n.d.-b. ‘Scikit Learn Documentation: PCA’. Scikit-Learn. Accessed 30 November 2024. https://scikit-learn/stable/modules/generated/sklearn.decomposition.PCA.html.

———. n.d.-c. ‘Scikit-Learn Documentation: K-Means’. Scikit-Learn. Accessed 8 November 2024. https://scikit-learn/stable/modules/clustering.html.

———. n.d.-d. ‘Scikit-Learn Documnetation: GaussianMixture’. Scikit-Learn. Accessed 30 November 2024. https://scikit-learn/stable/modules/generated/sklearn.mixture.GaussianMixture.html.

———. n.d.-e. ‘Scitkit-Learn Documentation: Hierarchical Clustering’. Scikit-Learn. Accessed 8 November 2024. https://scikit-learn/stable/modules/clustering.html.

SEC. 2017. ‘Investor Alert: Ponzi Schemes Using Virtual Currencies’. https://www.sec.gov/files/ia_virtualcurrencies.pdf.

Sergeenkov, Andrey. 2024. ‘What Is Ripple (XRP)?’ Forbes. 27 October 2024. https://www.forbes.com/sites/digital-assets/article/what-is-ripple-xrp/.

Shannon, C. E. 1948. ‘A Mathematical Theory of Communication’. *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shayegan, Mohammad Javad. n.d. ‘A Collective Anomaly Detection Method Over Bitcoin Network’. https://arxiv.org/pdf/2107.00925.

Shayegan, Mohammad Javad, Hamid Reza Sabor, Mueen Uddin, and Chin-Ling Chen. 2022. ‘A Collective Anomaly Detection Technique to Detect Crypto Wallet Frauds on Bitcoin Network’. *Symmetry* 14 (2): 328. https://doi.org/10.3390/sym14020328.

Statista. n.d. ‘Number of Cryptocurrencies 2013-2024’. Statista. Accessed 30 November 2024. https://www.statista.com/statistics/863917/number-crypto-coins-tokens/.

Subashi, Roland. 2024. ‘Cryptocurrencies and Money Laundering’. *Balkan Journal of Interdisciplinary Research* 10 (1): 55–62. https://doi.org/10.2478/bjir-2024-0005.

Sun Yin, Hao Hua, Klaus Langenheldt, Mikkel Harlev, Raghava Rao Mukkamala, and Ravi Vatrapu. 2019. ‘Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain’. *Journal of Management Information Systems* 36 (1): 37–73. https://doi.org/10.1080/07421222.2018.1550550.

Toyoda, Kentaroh, Tomoaki Ohtsuki, and P. Takis Mathiopoulos. 2018. ‘Multi-Class Bitcoin-Enabled Service Identification Based on Transaction History Summarization’. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 1153–60.

https://doi.org/10.1109/Cybermatics_2018.2018.00208.

Ul Hassan, Muneeb, Mubashir Husain Rehmani, and Jinjun Chen. 2023. ‘Anomaly Detection in Blockchain Networks: A Comprehensive Survey’. *IEEE Communications Surveys & Tutorials* 25 (1): 289–318. https://doi.org/10.1109/COMST.2022.3205643.

United Nations. n.d. ‘Economic Agent’. UNTERM - The United Nations Terminology Database. Accessed 22 October 2024. https://unterm.un.org/unterm2/en/view/bc6e1185-6da5-49ec-b4c7-7bf612a49236.

Victor, Friedhelm. 2020. ‘Address Clustering Heuristics for Ethereum’. In *Financial Cryptography and Data Security*, edited by Joseph Bonneau and Nadia Heninger, 12059:617–33. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51280-4_33.

Victor, Friedhelm, and Andrea Marie Weintraud. 2021. ‘Detecting and Quantifying Wash Trading on Decentralized Cryptocurrency Exchanges’. arXiv. https://doi.org/10.48550/arXiv.2102.07001.

Vlahavas, George, Kostas Karasavvas, and Athena Vakali. 2024. ‘Unsupervised Clustering of Bitcoin Transactions’. *Financial Innovation* 10 (1): 25. https://doi.org/10.1186/s40854-023-00525-y.

Wind, Wietse. 2024. ‘WietseWind/Fetch-Xrpl-Transactions’. https://github.com/WietseWind/fetch-xrpl-transactions.

Wu, Jiajing, Jieli Liu, Yijing Zhao, and Zibin Zheng. 2021. ‘Analysis of Cryptocurrency Transactions from a Network Perspective: An Overview’. *Journal of Network and Computer Applications* 190 (September):103139. https://doi.org/10.1016/j.jnca.2021.103139.

Yu, Congcong, Chen Yang, Zheng Che, and Liehuang Zhu. 2023. ‘Robust Clustering of Ethereum Transactions Using Time Leakage from Fixed Nodes’. *Blockchain: Research and Applications* 4 (1): 100112. https://doi.org/10.1016/j.bcra.2022.100112.

APPENDIX

Study	Data	Description
Reid and Harrigan (2011)	An Analysis of Anonymity in the Bitcoin System	**Approach:** Analyzed Bitcoin's transaction history to study anonymity through network structures. **Methodology:** Combined on-chain transaction analysis (topological network structure) with off-chain data (context discovery and flow analysis) to link addresses to real-world identities. **Key Insights:** Demonstrated that Bitcoin's pseudo-anonymity can be compromised by combining transaction patterns with external information, as shown in a case study investigating Bitcoin theft.
Androulaki et al. (2013)	Evaluating User Privacy in Bitcoin	**Approach:** Heuristic clustering and re-identification attacks applied to Bitcoin. **Methodology:** Used transaction features such as timing, sender/receiver indices, and transaction amounts to cluster users and evaluate privacy measures. **Key Insights:** Demonstrated that nearly 40% of user profiles could be recovered, showing significant privacy vulnerabilities even when recommended privacy measures are applied.
Meiklejohn et al. (2013)	A fistful of bitcoins: characterizing payments among men with no names	**Approach:** Heuristic clustering and re-identification attacks applied to the Bitcoin blockchain. **Methodology:** Grouped Bitcoin wallets using evidence of shared authority (e.g., multi-input heuristic) and empirically linked clusters to real-world entities through purchasing experiments. **Key Insights:** Demonstrated the visibility of Bitcoin's transaction flows despite pseudo-anonymity.
Nick (2015)	Data-Driven De-Anonymization in Bitcoin	**Approach:** Analyzed clustering strategies to group Bitcoin addresses belonging to the same wallet. **Methodology:** Used ground truth data from a Connection Bloom Filtering vulnerability. Evaluated clustering techniques for well-known multi-input heuristic and two newly proposed heuristics, combining them for improved results. **Key Insights:** Demonstrated that even modern Bitcoin wallets fail to fully protect users; the multi-input heuristic alone reveals 68.59% of addresses on average, with further improvements achieved through advanced heuristics.
Monaco (2015)	Identifying Bitcoin users by transaction behavior	**Approach:** Long-term transactional behavior analysis applied to Bitcoin to identify and verify account holders. **Methodology:** Analyzed transaction features such as timestamps, coin flow, transaction intervals, and network connectivity using a dynamical systems approach on users with 100–1000 monthly transactions for at least six months. **Key Insights:** Exploiting patterns in long-term transactional behavior reduces anonymity, revealing identifiable trends across outgoing and incoming transactions.
Harrigan and Fretter (2016)	The Unreasonable Effectiveness of Address Clustering	**Approach:** Address clustering using transaction micro-structures in the Bitcoin system. **Methodology:** Analyzed clustering heuristics using transaction micro-structures to analyze cluster growth and centrality.

		Key Insights: Identified key factors driving clustering effectiveness, including address reuse, incremental cluster growth, and super-clusters with high centrality.
Ermilov, Panov, and Yanovich (2017)	Automatic Bitcoin Address Clustering	**Approach:** Combined blockchain-based heuristics and off-chain data for Bitcoin address clustering. **Methodology:** Integrated common behavior patterns (e.g., common spending and one-time change) with off-chain information as additional "votes" for address separation to improve clustering precision. **Key Insights:** Demonstrated that combining blockchain and off-chain data enhances clustering accuracy, enabling more advanced de-anonymization and helping identify insecure Bitcoin usage patterns.
Jourdan et al. (2018)	Characterizing Entities in the Bitcoin Blockchain	**Approach:** Analyzed user-identifying patterns on the Bitcoin network to demonstrate unintended exposure of user information. **Methodology:** Investigated transaction patterns and network activity to uncover identifying information, highlighting widely available data-mining techniques. **Key Insights:** Bitcoin's pseudonymity can be compromised through transaction behaviors, as patterns and surrounding activity can reveal user identities.
Toyoda, Ohtsuki, and Mathiopoulos (2018)	Multi-Class Bitcoin-Enabled Service Identification Based on Transaction History Summarization	**Approach:** Multi-class service identification in Bitcoin using transaction history summarization from 2009-2017. **Methodology:** Extracted transaction history features from Bitcoin addresses and classified them into seven service categories using a supervised machine learning model. **Key Insights:** Achieved 72% accuracy in identifying services such as exchanges, gambling, mixers, and scams.
Victor (2020)	Address Clustering Heuristics for Ethereum	**Approach:** Developed address clustering heuristics tailored to Ethereum's account-based model, addressing limitations of UTXO-based methods. **Methodology:** Proposed heuristics based on deposit address reuse, multiple airdrop participation, and token authorization patterns, applied over 4 years of Ethereum data. **Key Insights:** Identified 17.9% of active externally owned addresses as clusters, representing over 340,000 entities, with the deposit address heuristic being the most effective.
Wu et al. (2021)	Analysis of cryptocurrency transactions from a network perspective: An overview	**Approach:** Comprehensive review of existing de-anonymization methods on different blockchains **Key Insights:** De-anonymization methods are grouped into three main types: transaction property-based, behavior-based, and off-chain information-based approaches.

Table 16: Overview of key work of heuristics

	Authors	Results
Supervised Agent Classification	*Lin et al. 2019*	Utilized a labeled dataset of 26,313 BTC addresses categorized into six groups to train several classification models. They achieved a Micro-F1 score of 0.87 and a Macro-F1 score of 0.86 using LightGBM.
	Harlev et al. 2018	Worked with a labeled dataset of 434 BTC addresses categorized into ten groups to train classification models. They achieved an accuracy of 77% and an F1 score of 75% using Gradient Boosting.
Unsupervised Agent Segmentation/Clustering	*Payette, Schwager, and Murphy 2017*	Used several unsupervised models to cluster 250,000 Ethereum addresses. Found K-Means as their best performing model using the Calinski-Harabasz score as a metric and identified 4 clusters. They didn't link the clusters to real word categories.
	Birrane 2017	Clustered 31,006 Ethereum users using K-Means into 15 clusters. Used the Elbow method to determine the optimal number of clusters. 99,59% of the records were grouped in one cluster. The clusters were not linked to real world categories.
	Ermilov, Panov, and Yanovich 2018	They used a probabilistic model and off-chain data to cluster around 95 million BTC addresses into 6 groups. No metrics that test cluster separation have been used.
	Vlahava, Karasavvas, and Vakali 2024	Clustered transactions (instead of addresses) into five clusters using trimmed K-Means achieving a Silhouette score of 0.78. They used off-chain data to verify their results. Gambling, exchanges, and services were almost completely grouped together.

Table 17: Overview of key work of agent classification

Study	Data	Methods	Key Results
Ostapowicz & Żbikowski (2019)	ETH (2,200 fraudulent wallets, 349,999 non-fraudulent wallets)	Random Forest , SVM, XGBoost	Random Forest: Best recall (84.92%), FPR (9.69%); XGBoost: similar to Random Forest, RF slightly better; SVM: High recall, but high FPR (less practical).
Farrugia et al. (2020)	ETH network (4,681 accounts, 2,179 flagged, 2,502 normal accounts)	XGBoost	Best accuracy: 0.96, F1 score: 0.96. Generalization and resilience towards overfitting.
Lorenz et al. (2021)	BTC transactional dataset (203,769 transactions), comparing supervised vs. unsupervised models	Supervised: Random Forest Unsupervised: KNN, PCA, Isolation Forest	Unsupervised models performed below supervised baseline. Fraudulent transactions not detected as outliers.
Nerurkar et al. (2021)	BTC (non-binary classification task: darknet markets, exchanges, gambling, Ponzi, unclassified)	Log Regression, SVM, Random Forest, XGBoost	Random Forest: best accuracy (0.92)

Table 18: Overview of key work of fraud detection

Feature	Description
payment_with_xrp_ratio	Ratio of payment transactions with a specified XRP amount to total transactions.
payment_without_xrp_ratio	Ratio of payment transactions without a specified XRP amount to total transactions.
offer_create_with_xrp_ratio	Ratio of offercreate transactions with a specified XRP amount (TakerGets/TakerPays) to total transactions.
offer_create_without_xrp_ratio	Ratio of offercreate transactions without a specified XRP amount to total transactions.
offer_cancel_ratio	Ratio of offercancel transactions to total transactions.
trust_set_ratio	Ratio of trustset transactions to total transactions.
escrow_ratio	Ratio of escrow transactions (all escrow-related types) to total transactions.
nftoken_create_offer_ratio	Ratio of nftokencreateoffer transactions to total transactions.
nftoken_cancel_offer_ratio	Ratio of nftokencanceloffer transactions to total transactions.
nftoken_accept_offer_ratio	Ratio of nftokenacceptoffer transactions to total transactions.
nftoken_mint_ratio	Ratio of nftokenmint transactions to total transactions.
total_transactions	Count of all transactions (incoming and outgoing) involving the account.
average_sent_amount_xrp	Average XRP amount sent in payment transactions where the address is the sender.
average_received_amount_xrp	Average XRP amount received in payment transactions where the address is the receiver.
median_sent_amount_xrp	Median XRP amount sent in payment transactions where the address is the sender.
median_received_amount_xrp	Median XRP amount received in payment transactions where the address is the receiver.
stddev_sent_amount_xrp	Standard deviation of XRP amounts sent in payment transactions where the address is the sender.

stddev_received_amount_xrp	Standard deviation of XRP amounts received in payment transactions where the address is the receiver.
payment_as_account_ratio	Ratio of payment transactions where the address is the sender (account) to total transactions.
payment_as_destination_ratio	Ratio of payment transactions where the address is the receiver (destination) to total transactions.
source_tag_present_ratio	Ratio of transactions with a non-null SourceTag when the address is the sender to total transactions.
destination_tag_present_ratio	Ratio of transactions with a non-null DestinationTag when the address is the receiver to total transactions.
unique_transaction_partner_ratio	Ratio of unique transaction partners in payment transactions to total transactions.
payment_small_amounts_sender_ratio	Ratio of payment transactions under specific amounts (10, 20, 50, 100, 200, 300, 500 XRP) sent to total transactions.
payment_small_amounts_receiver_ratio	Ratio of payment transactions under specific amounts (10, 20, 50, 100, 200, 300, 500 XRP) received to total transactions.
payment_large_amounts_sender_ratio	Ratio of payment transactions over specific amounts (500, 1,000, 10,000, 100,000, 1,000,000 XRP) sent to total transactions.
payment_large_amounts_receiver_ratio	Ratio of payment transactions over specific amounts (500, 1,000, 10,000, 100,000, 1,000,000 XRP) received to total transactions.
absolute_active_period	Difference in days between the first and last transaction timestamps for the account, indicating the span of activity.
active_days	Count of unique days during which the account conducted at least one transaction.
transaction_frequency	Number of transactions over the active period of the account.
unique_destination_partners	Count of unique transaction partners where the address is the sender.
unique_account_partners	Count of unique transaction partners where the address is the receiver.
transaction_ratio_20_largest_exchanges	Ratio of transactions associated with the 20 largest exchanges to total transactions.
total_sum_sent	Sum of all outgoing transaction values for the account over the observed period.
total_sum_received	Sum of all incoming transaction values for the account over the observed period.

Table 19: Account level feature set

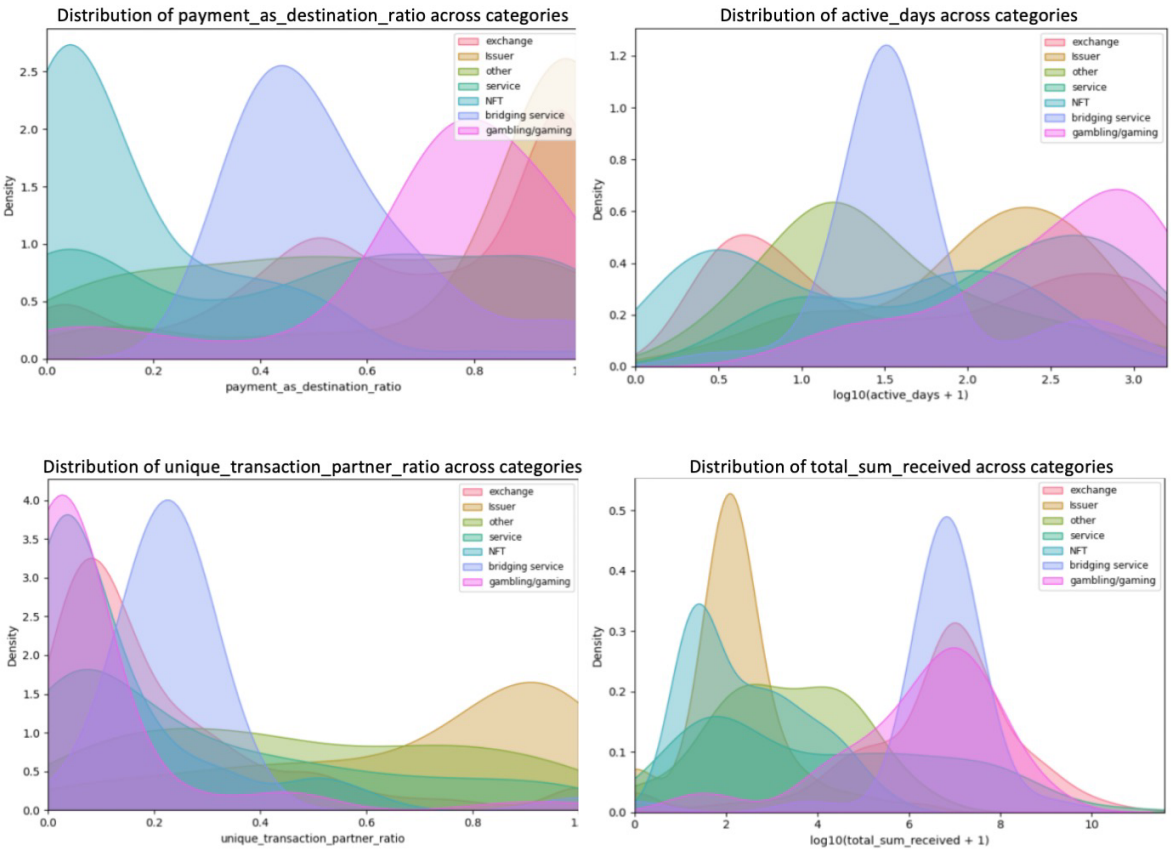

Figure 3: Distribution of various features across categories

Feature	H-statistic	p-value
nftoken_mint_ratio	1285.589967	1.43E-274
payment_without_XRP_ratio	788.167695	5.55E-167
payment_over_1000_xrp_receiver_ratio	630.5956361	5.85E-133
nftoken_create_offer_ratio	612.526185	4.63E-129
payment_over_10000_xrp_receiver_ratio	560.8910019	6.34E-118
nftoken_cancel_offer_ratio	538.617142	4.01E-113
nftoken_accept_offer_ratio	480.5705936	1.29E-100
total_sum_received	470.6342545	1.77E-98
payment_over_1000_xrp_sender_ratio	470.1915373	2.21E-98
payment_over_500_xrp_sender_ratio	443.8323629	1.04E-92
payment_over_10000_xrp_sender_ratio	437.7274626	2.15E-91
trust_set_ratio	410.7263861	1.38E-85
payment_over_100000_xrp_receiver_ratio	402.3527507	8.72E-84
total_sum_sent	394.8461681	3.58E-82
destination_tag_present_ratio	347.7990786	4.58E-72
payment_over_100000_xrp_sender_ratio	335.6828905	1.83E-69
payment_over_1000000_xrp_sender_ratio	287.5388858	3.82E-59
offer_create_with_XRP_ratio	257.1853796	1.19E-52
unique_transaction_partner_ratio	235.2377555	5.84E-48
payment_over_1000000_xrp_receiver_ratio	177.5972987	1.10E-35
unique_account_partners	162.461934	1.78E-32
offer_cancel_ratio	162.419805	1.82E-32
payment_under_500_xrp_receiver_ratio	136.9778122	4.35E-27
ratio_known_counterparty_transactions	136.1662466	6.45E-27
total_transactions	114.1729806	2.72E-22
unique_destination_partners	110.2187033	1.83E-21
source_tag_present_ratio	107.8585669	5.72E-21
active_days	103.4768496	4.72E-20
payment_under_500_xrp_sender_ratio	40.40788751	3.79E-07
offer_create_without_XRP_ratio	36.59046804	2.12E-06
escrow_ratio	14.77972275	0.022040934
median_sent_amount_xrp	NaN	NaN
median_received_amount_xrp	NaN	NaN
stddev_received_amount_xrp	NaN	NaN

Table 20: Kruskal-Wallis test results

Feature	NFT	Exchange	Gambling	Bridging Service	Issuer	Others	Service
absolute_active_period	-0.657	-0.046	-1.026	-0.122	0.061	0.060	0.124
active_days	0.294	-0.146	0.815	-0.038	0.008	-0.066	0.044
average_received_amount_xrp	-0.136	0.084	-0.406	0.003	-0.030	-0.016	-0.025
destination_tag_present_ratio	-1.338	0.213	1.001	-0.141	0.035	-0.103	-0.210
escrow_ratio	-0.220	-0.086		-0.011	-0.024	0.042	0.030
median_received_amount_xrp	-0.154	0.058	-0.949	-0.004	-0.028	-0.007	-0.009
nftoken_accept_offer_ratio	0.084	-0.057		-0.016	-0.152	0.038	-0.001
nftoken_cancel_offer_ratio	-0.030	0.003		-0.014	0.071	0.008	0.002
nftoken_create_offer_ratio	0.491	-0.102		-0.026	-0.050	-0.091	-0.022
nftoken_mint_ratio	0.844	-0.102		-0.044	-0.050	-0.104	-0.080
offer_cancel_ratio	0.083	-0.030		-0.039	0.054	0.031	0.005
offer_create_with_XRP_ratio	-0.049	-0.119		-0.045	-0.029	0.079	0.111
offer_create_without_XRP_ratio	-0.114	-0.019		-0.018	-0.020	-0.034	0.056
payment_over_1000000_xrp_receiver_ratio	-0.268	0.167	-1.109	-0.016	-0.049	-0.027	-0.033
payment_over_1000000_xrp_sender_ratio	-0.503	0.229	-2.319	-0.035	-0.081	-0.037	-0.042
payment_over_100000_xrp_sender_ratio	-0.685	0.193	-2.657	-0.036	-0.089	-0.042	-0.042
payment_under_10_xrp_sender_ratio	0.011	-0.016	0.304	0.054	0.006	0.091	-0.081
ratio_known_counterparty_transactions	-1.290	0.265	0.093	-0.040	-0.201	-0.050	-0.099
source_tag_present_ratio	-0.132	-0.005	-0.110	-0.039	-0.089	-0.010	0.078
stddev_received_amount_xrp	0.020	-0.064	-0.207	0.000	-0.029	-0.016	0.055
total_sum_received	-0.069	0.023	-0.462	0.000	-0.031	0.010	0.000
total_sum_sent	-0.118	0.117	-0.957	0.010	-0.042	0.017	-0.052
total_transactions	-0.389	-0.017	0.087	-0.010	-0.054	0.014	-0.040
transaction_frequency	-0.549	-0.020	0.034	-0.023	-0.055	-0.024	0.039
trust_set_ratio	0.157	-0.218		-0.029	0.058	0.052	0.115
unique_account_partners	-0.007	0.065	-0.355	0.020	-0.013	0.020	-0.027
unique_destination_partners	0.129	0.059	-0.118	-0.028	-0.071	-0.028	0.028
unique_transaction_partner_ratio	-0.268	-0.168	-0.976	-0.045	0.399	0.162	0.066

Table 21: Logistic regression results with features listed vertically and coefficients shown horizontally

Feature	NFT	Exchange	Gambling	Bridging Service	Issuer	Others	Service
absolute_active_period	0.518	0.955	0.358	0.885	1.063	1.062	1.132
active_days	1.342	0.864	2.259	0.963	1.009	0.936	1.046
average_received_amount_xrp	0.873	1.088	0.666	1.003	0.970	0.984	0.976
destination_tag_present_ratio	0.262	1.237	2.721	0.869	1.036	0.902	0.810
escrow_ratio	0.803	0.918	1.000	0.989	0.976	1.043	1.031
median_received_amount_xrp	0.857	1.059	0.387	0.996	0.972	0.993	0.991
nftoken_accept_offer_ratio	1.087	0.944		0.984	0.859	1.038	0.999
nftoken_cancel_offer_ratio	0.970	1.003		0.986	1.074	1.008	1.002
nftoken_create_offer_ratio	1.635	0.903		0.974	0.952	0.913	0.978
nftoken_mint_ratio	2.326	0.903		0.957	0.951	0.901	0.923
offer_cancel_ratio	1.087	0.970		0.962	1.056	1.032	1.005
offer_create_with_XRP_ratio	0.952	0.888		0.956	0.971	1.082	1.118
offer_create_without_XRP_ratio	0.892	0.981		0.982	0.980	0.967	1.057
payment_over_1000000_xrp_receiver_ratio	0.765	1.181	0.330	0.985	0.952	0.974	0.967
payment_over_1000000_xrp_sender_ratio	0.605	1.258	0.098	0.965	0.922	0.963	0.959
payment_over_100000_xrp_sender_ratio	0.504	1.213	0.070	0.965	0.915	0.959	0.959
payment_under_10_xrp_sender_ratio	1.011	0.984	1.356	1.056	1.006	1.095	0.922
ratio_known_counterparty_transactions	0.275	1.304	1.098	0.961	0.818	0.951	0.906
source_tag_present_ratio	0.876	0.995	0.896	0.962	0.914	0.990	1.081
stddev_received_amount_xrp	1.020	0.938	0.813	1.000	0.971	0.984	1.056
total_sum_received	0.934	1.023	0.630	1.000	0.969	1.010	1.000
total_sum_sent	0.889	1.124	0.384	1.011	0.959	1.017	0.949
total_transactions	0.678	0.983	1.091	0.990	0.947	1.014	0.961
transaction_frequency	0.577	0.980	1.034	0.977	0.946	0.977	1.039
trust_set_ratio	1.170	0.805	1.000	0.972	1.060	1.053	1.122
unique_account_partners	0.993	1.067	0.701	1.021	0.987	1.021	0.973
unique_destination_partners	1.137	1.061	0.889	0.973	0.931	0.972	1.029
unique_transaction_partner_ratio	0.765	0.845	0.377	0.956	1.490	1.176	1.068

Table 22: Logistic regression results with features listed vertically and odds ratios shown horizontally

Features for K-means and Gaussian Mixture
offer_create_with_xrp_ratio
offer_create_without_xrp_ratio
offer_cancel_ratio
trust_set_ratio
escrow_ratio
nftoken_create_offer_ratio
nftoken_cancel_offer_ratio
nftoken_accept_offer_ratio
nftoken_mint_ratio
total_transactions
average_sent_amount_xrp
average_received_amount_xrp
median_sent_amount_xrp
median_received_amount_xrp
stddev_sent_amount_xrp
stddev_received_amount_xrp
unique_transaction_partner_ratio
payment_small_amounts_sender_ratio
payment_small_amounts_receiver_ratio
payment_large_amounts_sender_ratio
payment_large_amounts_receiver_ratio
absolute_active_period
active_days
transaction_frequency
unique_destination_partners
unique_account_partners
transaction_ratio_20_largest_exchanges
total_sum_received

Table 23: Features used for K-Means clustering and Gaussian Mixture

Features for Agglomerative clsutering
offer_create_with_xrp_ratio
offer_create_without_xrp_ratio
offer_cancel_ratio
trust_set_ratio
escrow_ratio
nftoken_create_offer_ratio
nftoken_cancel_offer_ratio
nftoken_accept_offer_ratio
nftoken_mint_ratio
total_transactions

average_sent_amount_xrp
average_received_amount_xrp
median_sent_amount_xrp
median_received_amount_xrp
stddev_sent_amount_xrp
stddev_received_amount_xrp
unique_transaction_partner_ratio
payment_small_amounts_sender_ratio
payment_small_amounts_receiver_ratio
payment_large_amounts_sender_ratio
payment_large_amounts_receiver_ratio
absolute_active_period
active_days
transaction_frequency
unique_destination_partners
unique_account_partners
transaction_ratio_20_largest_exchanges

Table 24: Features used for Agglomerative clustering

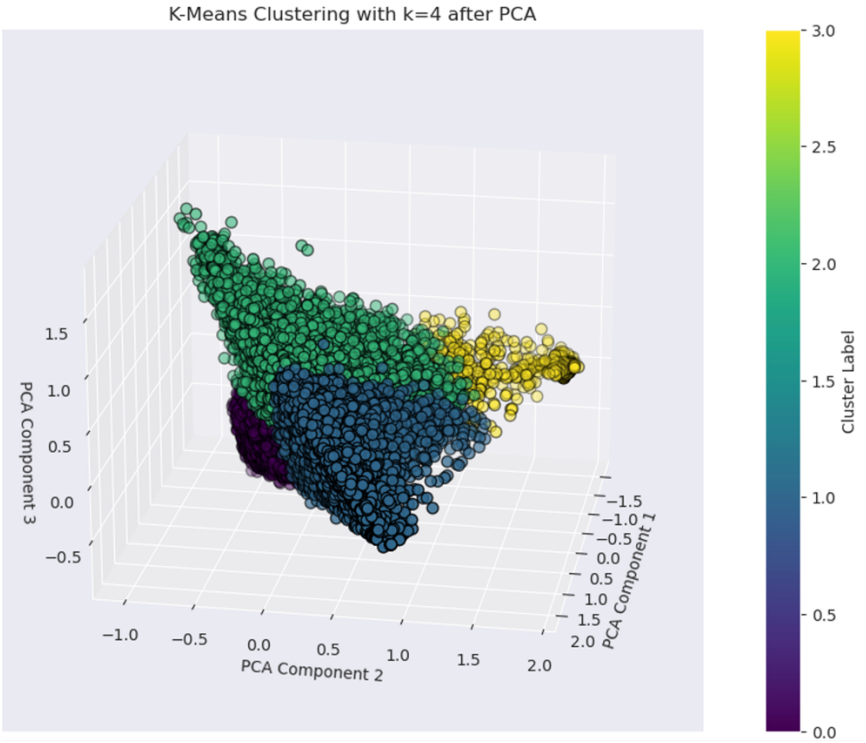

Figure 4: Visualization of the three principal components of K-Means with k=4

Metric	Agglomerative Clustering with PCA (k=5)	Gaussian Mixture with PCA (k=6)	K-Means with PCA (k=4)
Running Time (s)	452.97	6.16	5.67
Memory Usage (MB)	9.13	30.11	15.82

Table 25: Comparison of runtime and memory usage of the unsupervised models

Features for LightGBM
absolute_active_period
active_days
average_received_amount_xrp
average_sent_amount_xrp
destination_tag_present_ratio
median_received_amount_xrp
median_sent_amount_xrp
nitoken_create_offer_ratio
offer_cancel_ratio
offer_create_with_XRP_ratio
offer_create_without_XRP_ratio
payment_as_account_ratio
payment_over_100000_xrp_receiver_ratio
payment_over_100000_xrp_sender_ratio
payment_over_1000_xrp_receiver_ratio
payment_over_1000_xrp_sender_ratio
payment_over_10000_xrp_receiver_ratio
payment_over_10000_xrp_sender_ratio
payment_over_500_xrp_receiver_ratio
payment_over_500_xrp_sender_ratio
payment_under_10_xrp_sender_ratio
payment_under_100_xrp_receiver_ratio
payment_under_20_xrp_receiver_ratio
payment_under_200_xrp_receiver_ratio
payment_under_200_xrp_sender_ratio
payment_under_50_xrp_sender_ratio
payment_under_500_xrp_receiver_ratio
payment_without_XRP_ratio
ratio_known_counterparty_transactions
source_tag_present_ratio
stddev_received_amount_xrp
total_sum_received
total_sum_sent
transaction_frequency
trust_set_ratio
unique_account_partners

unique_destination_partners
unique_transaction_partner_ratio

Table 26: Final selection of features for LightGBM classifier

Metric	Best Random Forest	Best LightGBM
Training Time (s)	49.88	1.03
Inference Time per Instance (ms)	0.050	0.002
Peak Memory Usage (MB)	458.30	491.61

Table 27: Comparison of computational performance

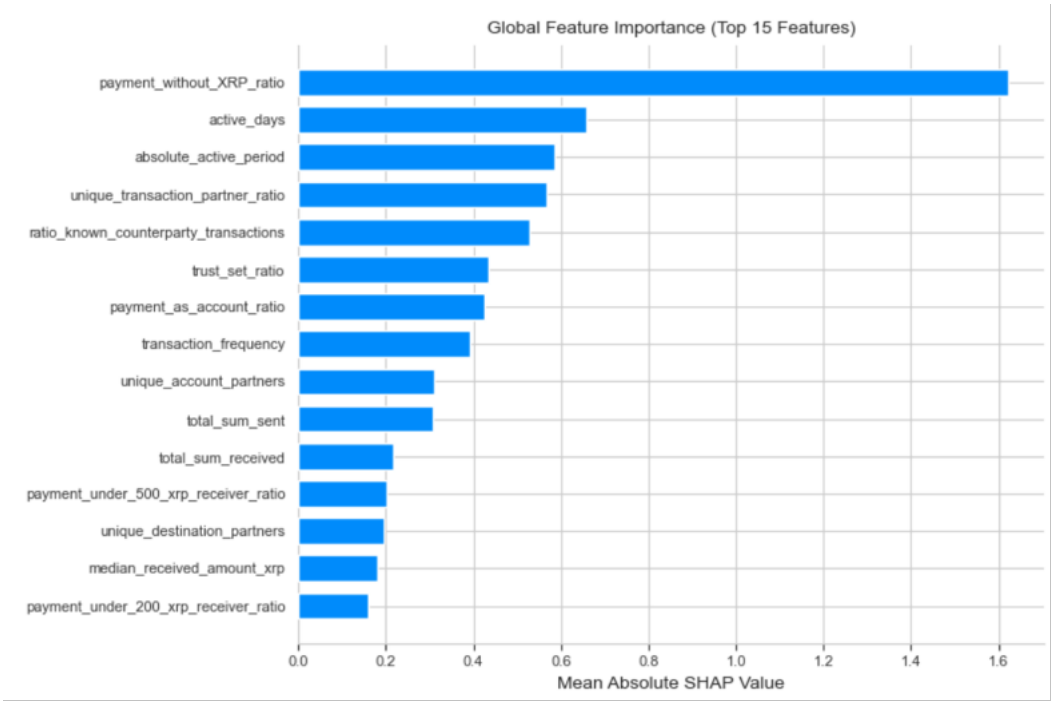

Figure 5: Global feature importance top 15 – LightGBM