Laundromats: More Than Just Missing Socks

Improving the estimation of the False Negative Rate of money laundering detection at Dutch banks

D.G. van de Pol

TU Delft

KPMG

MSc Thesis in Engineering & Policy Analysis

Faculty of Technology, Policy, and Management

Laundromats: More Than Just Missing Socks

Improving the estimation of the False Negative Rate of money laundering detection at Dutch banks

by

D.G. van de Pol

Student number: 4678834

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on December 14, 2023.

Graduation committee

Dr. ir. Maarten Kroesen	TU Delft
Dr. Savvas Zannettou	TU Delft, supervisor
Esther Kuikman MSc	KPMG, supervisor

Project duration: June 2023 – December 2023

Preface

Recalling the days when 'smurfing' was merely a cartoon activity, not a money laundering technique, my thesis journey began six months ago as a Graduate Intern at KPMG Forensic Technology. Wanting to 'do something for the greater good' using a quantitative machine learning approach, I felt privileged to have the option to work on the topic of Anti-Money Laundering (AML) at a renowned consultancy firm while also enjoying the tremendous support and resources TU Delft has to offer. Getting to know the AML domain was challenging, but a great adventure, nevertheless.

Thus, with great pride, I present my master thesis titled 'Laundromats: More Than Just Missing Socks: Enhancing the Estimation of the False Negative Rate in Money Laundering Detection at Dutch Banks'. This study is conducted in partial fulfillment of the requirements for the degree of Master of Science in Engineering & Policy Analysis at the Faculty of Technology, Policy, and Management at Delft University of Technology.

This thesis aims to assess the feasibility of estimating the False Negative Rate of unreviewed transactions of current rule-based AML Transaction Monitoring systems in place at Dutch banks. This knowledge holds value for all actors involved in the Dutch Anti-Money Laundering landscape, more particularly banks and regulators.

I wish to express my gratitude to my first supervisor, Dr. Savvas Zannettou, for all of his enthusiasm and invaluable insights over the past six months. Our weekly Friday morning discussions were a source of continual inspiration, invariably leading to weekends filled with new plans and ideas. In addition, I fully got the freedom to choose and explore my personally preferred topic, for which I am grateful. Special thanks also go to my second supervisor, Dr. ir. Maarten Kroesen, for his constructive feedback as second reader and thesis committee member, as well as for graciously agreeing to chair my graduation committee at short notice.

Third, I am profoundly thankful to my whole team at KPMG, especially my supervisors Joyce Pebesma and Esther Kuikman, with whom our weekly meetings were both enlightening and enjoyable. I also would like to express my gratitude to the interviewees for their invaluable contributions and insights. Finally, my deepest appreciation goes to my girlfriend, parents and all those who have supported me, offering their time for brainstorming, discussion, and review, and for providing unwavering motivation during challenging times.

This work represents the culmination of my academic journey at TU Delft, and it is with a sense of accomplishment and pride that I share it with you. I hope you find reading this thesis as enriching and enjoyable as I found writing it.

Dani van de Pol

Utrecht, November 2023

Executive Summary

The Dutch banking sector is mandated to identify and report transactions that may signify money laundering (ML) activities. Banks have been reliant on rule-based transaction monitoring (TM) systems that flag transactions exceeding predefined thresholds. While such systems are instrumental in filtering potential ML transactions, the inherently small prevalence rate of ML occurring in the vast majority of financial transactions causes these systems to produce only a limited number of flagged transactions. Furthermore, as flagged transactions are only those surpassing certain rule thresholds, the alerts are biased toward presumed risk distributions. Consequently, this causes the performance regarding transactions that go unreviewed but should have been flagged, so-called false negatives, to remain unknown. This lack of understanding is a critical gap in the efficacy of current anti-money laundering (AML) controls and motivates the need for better insights into this False Negative Rate (FNR).

Addressing this critical need for enhanced discernment of FNR, this study aims to improve this knowledge gap by answering the following research question: 'To what extent could a supervised machine learning classifier, when trained on historical alerts, assist Dutch banks in estimating the False Negative Rate of rule-based transaction monitoring systems concerning unreviewed transactions?'. To achieve this goal, this study adopts a mixed-methods research design by combining a literature review with seven interviews with domain specialists to acquire insights into transactional ML typologies and the underlying indicators and thresholds employed in existing rule-based TM systems. The study further extended to the development of eight different types of supervised machine learning classifiers, applied both using their default settings as well as with balancing measures in place when possible. These classifiers were trained on two synthetically generated datasets of both 180 million transactions, one with a high and one with a low ML prevalence rate, indicating the relative frequency of ML transactions. These mirrored real transactional patterns in order to evaluate the feasibility of estimating the FNR pertaining to unreviewed transactions utilizing historical data on flagged transactions. In addition to establishing the effect of both different ML prevalence rates on performance, we also explored whether combining data from multiple financial institutions into one shared information perspective could be of additional value.

The findings indicate that the classifiers struggle to accurately predict the FNR, especially in scenarios of low ML prevalence and without combining information from multiple institutions. There is a significant discrepancy between the actual (0.729 to 0.988) and predicted FNRs (0.156 to 0.649), even in higher ML prevalence settings. The low performance, as evidenced by poor Area Under Precision-Recall Curve (AUPRC) and Matthews Correlation Coefficient (MCC) scores, highlights the challenges in using machine learning for ML transaction detection in Dutch banking, calling for further research and development of more advanced detection models.

Despite the restrained success in accurately estimating FNR through supervised machine learning classifiers, the insights derived from this research are of considerable value. They prompt a critical examination of the current TM systems and suggest a pivot toward more sophisticated machine learning techniques for FNR estimation. The findings serve as a start for Dutch banks to refine their AML strategies and enhance the integrity of financial systems by exploring and potentially integrating more nuanced and data-driven approaches for detecting ML activities.

To address limitations such as modest prediction accuracy and a significant gap between actual and predicted FNRs, it is essential to explore typology-specific models that are finely tuned to distinct ML patterns. This recommendation aligns with the key limitation of current classifiers' performance being intertwined with various parameters and typologies. Collaborative data sharing among financial institutions, underpinned by secure protocols, also emerges as a crucial

strategy for enhancing ML detection, overcoming the challenge of limited data representativeness.

Future research should focus on developing balanced datasets with advanced sampling techniques to better address class imbalance, and exploring graph-based machine learning methods that capture complex transactional relationships. Additionally, considering the limitation of current datasets validated primarily against U.S. data, further research must ensure contextual relevance to the Dutch banking landscape. This includes the integration of external data sources for a broader contextual understanding and long-term data analysis for evolving ML tactics.

Table of Contents

List of Tables

List of Figures

Abbreviation	Definition
AML	Anti-Money Laundering
ATL	Above-The-Line
AUPRC	Area Under Precision-Recall Curve
CDD	Customer Due Diligence
CTF	Counter-Terrorist Financing
ETP	Expected Transaction Profile
FIU	Financial Intelligence Unit
FNR	False Negative Rate
KYC	Know Your Customer
MCC	Matthews Correlation Coefficient
ML	Money Laundering
TM	Transaction Monitoring
Wwft	Wet Ter voorkoming van Witwassen en Financieren van Terrorisme (Anti-Money Laundering and Anti-Terrorist Financing Act)

1. Introduction

Money laundering (ML) is a significant global issue, and the Netherlands plays a prominent role on the world stage with an estimated amount of 16 billion euros being laundered annually (Baazil, 2022; Teeffelen, 2018). To counter this, gatekeepers in the Dutch financial system, such as banks, are required by the Anti-Money Laundering and Anti-Terrorist Financing Act (Wwft) to adopt anti-money laundering (AML) measures. This means that they must assess the risks customers entail (Know Your Customer; KYC) and report unusual transactions to the Financial Intelligence Unit (FIU) (DNB, 2021; FIU-Nederland, n.d.-b). Therefore, Dutch banks conduct thorough Customer Due Diligence (CDD) before onboarding new clients and continue to monitor their customers' activities while providing services, using, among others, so-called Transaction Monitoring (TM) processes. Banks increasingly deploy computer systems for those TM purposes, which create alerts for flagged transactions deemed unusual for subsequent manual investigation (Chau & Nemcsik, 2020). These systems may adopt either a traditional rule-based approach - e.g., flagging transactions that exceed a certain threshold – or a machine learning approach, which e.g., operates based on identified patterns in historical data. Currently, predominantly traditional rule-based TM systems are in place (Gerlings & Constantiou, 2022, p. 3474).

Interestingly enough, however, the predominant focus of academia within this domain has been directed towards research further enhancing and refining ML detection algorithms for TM purposes. Conversely, more traditional rule-based TM systems have received relatively limited attention for potential advancements. Thus, despite the prevalence of rule-based TM systems in current practice, a significant gap exists in the academic research concerning those systems in the context of their efficiency and potential vulnerabilities. This lack of interest persists whether the objective is to enhance these systems directly or to derive a more generalized understanding of their workings and limitations.

This lack of academic focus leads us to an important issue. Due to the selective approach of rule-based TM systems as well as to the inherently low prevalence rate of ML relative to total cashflows, TM systems within most (larger) banks only flag a relatively limited percentage of transactions (UNODC, n.d.-a). This limited number of alerts consists only of those transactions with a high deemed risk, thus exceeding preset thresholds. All alerts are then manually investigated after which they are reported as unusual or not. This gives banks insight into the unusual rate on a presumed high-risk subset of the transactions – namely those being flagged - but leaves a blind spot regarding the usually vast majority of transactions that never undergo manual investigation, creating a potential risk (Tertychnyi et al., 2020; Vassallo et al., 2021). Simply rephrased, banks possess awareness of the unusual rate of transactions they have detected but lack knowledge of the transactions they have missed. However, potentially quite a material share of the 'non-flagged' transactions might warrant manual review. Thus, there exists an urgent requirement to improve the evaluations of rates of TM models regarding transactions that have been inaccurately left unreviewed. This can be measured using the False Negative Rate (FNR), a metric that quantifies the rate at which actual unusual transactions are mistakenly not flagged by the system.

To address the knowledge gap, this study aims to assess the feasibility of estimating the FNR pertaining to unreviewed transactions. The objective is to train models on historical typology-specific subsets of transactional data which can then detect ML transactions for a larger unseen set. To this end, the methodology employs supervised machine learning classifiers as they are adept at learning from complex data patterns and adapting to new ML typologies, surpassing the limited scope of rule-based systems that rely on fixed thresholds. Due to the difficulty of obtaining (accurate) real-world transactional data and ensuring accurate ground-truth labeling, synthetic transaction datasets with labeled ML behavior replicating the complexities of real banking transactions were used within a clinical setting. Against this backdrop, this study sets forth the following research question:

To what extent could a supervised machine learning classifier, when trained on historical alerts, assist Dutch banks in estimating the False Negative Rate of rule-based transaction monitoring systems concerning unreviewed transactions?

Currently, the forefront of the field suggests the theoretical possibility of statistically estimating the FNR pertaining to unseen transactions. However, the concept remains significantly underexplored to date due to academia's preoccupation with enhancing machine learning approaches for AML purposes, thereby creating a lack of interest in rule-based TM systems.

However, interest regarding the feasibility of FNR estimation has increased due to multiple reasons. First, the New York State Department of Financial Services issued a new regulation in 2017, which imposed measures on regulated financial institutions to test whether their AML TM systems are on par with the regulation and the institutions' risk assessment (NYSDFS, n.d.; Ortwine, 2017). As it mandates the use of technology within AML model risk management, it has led to an increased interest in statistical evaluations of TM systems. Due to the importance of New York City as a worldwide financial hub, this regulation also impacted non-US international banks with branches or offices in the U.S., including Dutch ones, directly and indirectly (McGettigan, 2017). As the control over easily identifiable and observable risks continues to strengthen, attention is now shifting toward the identification and assessment of latent vulnerabilities within existing TM systems. Second, until recently, TM systems were predominantly overseen by the compliance departments of banks, which generally tend to consider them from a legal rather than a technical or statistical perspective.

This study is structured into two research phases and four sub-questions, aiming to guide the inquiry and subsequently answer the overarching research question. The deliverable of the first phase is an overview of the most common transactional ML typologies in the Netherlands, their key indicators and the thresholds. This informed the development of a simulated rule-based TM system, providing a historical alert dataset and the actual FNR of the rule-based TM system. The second phase outlines which statistical methods can be used to classify unreviewed ML transactions based on historically reviewed transactions in order to obtain a predicted FNR.

Phase 1 – Understanding Typologies and Rule-Based Monitoring

- SQ1. What are the most common transactional money laundering typologies within the Dutch banking system?
- SQ2. Which indicators are most predictive of common transactional money laundering typologies in the Dutch banking sector?
- SQ3. What are the thresholds used for key indicators in the rule-based transaction monitoring systems employed by Dutch banks?

Phase 2 – Machine Learning Classification for Transaction Analysis

- SQ4. Which supervised machine learning classifiers, recommended by literature, demonstrate the highest efficacy in predicting unusual transactions indicative of money laundering?

This study has been conducted in two phases. The first phase consisted of a literature review and seven interviews with AML specialists and entailed finding out how rule-based TM models of banks decide on flagging certain transactions. These qualitative methods were deliberately chosen for their ability to capture the nuanced perspectives and practical knowledge of practitioners and experts in the field. They provided valuable insights, revealing the decision-making processes and operational challenges inherent in these systems. Additionally, this phase outlined what common ML typologies in the Netherlands are and how the thresholds of key indicators for those typologies can be assumed. Altogether, this knowledge informed the development of a simulated rule-based TM system, providing artificially generated (synthetic) datasets of bank transactions containing multiple types of ML typologies common within the Dutch financial banking system.

In the second phase, we selected supervised machine learning classifiers known for their effectiveness in detecting unusual transactions. These classifiers were trained using typology-specific subsets of transactions that would have triggered real-world alerts by exceeding the predefined thresholds of ML indicators. By making predictions for unreviewed transactions using the trained classifiers, statistical evaluations of these predictions provided insights into the feasibility of estimating the FNR pertaining to unreviewed transactions.

This study aims to contribute to a better understanding of the current rule-based TM system's overall performance. Besides advancing academic knowledge, Dutch banks and regulatory authorities can leverage this information to more effectively assess risks associated with unreviewed transactions and to better align these risks with their risk appetites, promoting a more integer Dutch financial system. Moreover, the resolution of this research question holds significant societal implications and extends beyond the financial sector. Firstly, it aligns with Goal 16 of the United Nations Sustainable Development Goals, which entails the need for robust measures to detect and prevent ML as essential to fostering sustainable economic development (Canhoto, 2021; UNODC, n.d.-b). The societal significance of this issue is further underlined by ongoing initiatives from the European Union and the Dutch government. Both are intensifying their regulatory frameworks and operational efforts to counter financial crime, including ML (European Commission, 2021; Ministerie van Algemene Zaken, 2022). Should the outcomes prove promising, the concept of FNR estimation could pave the way for subsequent evaluations using real-world TM systems and data, employing methodologies such as above-the-line (ATL) testing.

Furthermore, this study seeks to scrutinize the possibilities of estimating the FNR of rule-based TM systems in identifying ML activities, specifically within large banks catering to retail customers. While AML is often mentioned together with Counter-Terrorist Financing (CTF) in legislation and public discourse, this study isolates its focus on TM designed for ML detection. Activities related to CTF, as well as other financial crimes such as bribery and corruption, are beyond the purview of this study. The rationale for emphasizing large banking institutions arises from their tendency to employ rule-based TM systems, owing to the sheer volume of transactions that exceed manual review capabilities. Smaller banks, which are generally specialized and have fewer transactions, thereby have less need for these automated systems and are thus excluded from this study. Given the diverse spectrum of banking clientele (e.g., large multinationals or private banking), this study narrows its focus to retail customers, which include both personal as well as business accounts. Rule-based TM systems are generally more effective in environments with high transaction volumes typically associated with retail banking. Given the significant divergence in ML typologies between retail and wholesale banking contexts, focusing on retail banking provides a more controlled environment for evaluating the efficacy of rule-based TM systems in estimating the FNR. Consequently, the scope of this study is confined to 'pure payment' transactions, excluding 'non-payment' transactions such as interest accruals or reinvestment activities.

The results conclusively demonstrate that supervised machine learning classifiers, when trained on historical alerts, are currently inadequate for assisting Dutch banks in accurately estimating the FNR of rule-based TM systems with respect to unreviewed transactions. Given the additional findings all banks' perspective models perform better, a critical recommendation and direction for future research lies in facilitating data sharing while maintaining privacy safeguards. Joint initiatives among banks have the potential to forge more resilient and inclusive models by pooling collective insights, all while ensuring compliance with privacy and data protection standards. Such collaborative endeavors are vital for crafting a more comprehensive and effective strategy for ML detection.

The remainder of this thesis is structured as follows. The following Chapter 2 provides the background and important related work of this study regarding ML typologies through transactions and how those can be detected using rule-based TM systems. In addition, statistical

methods for predicting ML transactions are discussed. This is followed by Chapter 3 which presents the methodology. Chapter 4 presents the insights gathered from the interviews, after which Chapter 5 outlines the creation of the used datasets and the experiments. Lastly, the results, discussion and conclusion are provided in Chapters 6, 7 and 8 respectively.

2. Background and Related Work

This section presents an overview of the current state-of-the-art literature within the fields of transactional ML typologies, key ML indicators and statistical methods for ML classification purposes to be able to estimate the FNR.

2.1 Money Laundering Definition

ML is generally understood as the illicit practice of disguising the origin of funds obtained through illegal activities (Levi & Reuter, 2006). The primary aim of this practice is to integrate these unlawfully obtained funds into the formal economy by obscuring their illegal origins and attributing a legal source to them (FinCEN, n.d.; Reuter & M. Truman, 2005). Article 420bis Sections 1 and 2 of the Dutch Criminal Code offer a legalistic definition, specifying that ML includes actions that intentionally obscure or misrepresent the authentic origin, source, or ownership of funds that are known to be directly or indirectly derived from criminal activities. Interpol (2023) amplifies this legal framework by encompassing any actions committed with the intent of obscuring or distorting the identity of illicitly acquired proceeds, thereby rendering them ostensibly lawful in origin. The various definitions, whether legal, academic, or institutional, converge on the essential characteristic of ML: the intent to legitimize illicitly obtained assets through obfuscation (Unger & Van Waarden, 2009). Transitioning to its historical context, let's delve into the evolutionary trajectory of this illicit activity.

2.2 History

Although the first ML occurred as early as 2000 BCE, it is commonly believed that the practice of ML as we understand it today soared during the U.S. prohibition period in the 1920s (Jass, KYC-Chain, 2019). This is when illegally obtained Mafia funds from alcohol imports found their way into the economy (UKALA, 2012). Nevertheless, the term 'money laundering' was not coined until much later, during the depiction of the Watergate scandal by The Guardian (Paxton, 2015; UKALA, 2012). AML, on the other hand, originated during the prohibition era as well with the first US AML law being put into place but has gotten more substantial only since 1970 when the US put more regulations in place in the war on drugs (Kenton, 2022). Ever since IT systems became mainstream, banks have used traditional fixed rule-based alert systems to flag unusual transactions (Eddin et al., 2022; Jullum et al., 2020). Moving from historical underpinnings to specific methods, the next section elaborates on various typologies of ML.

2.3 Money Laundering Typologies

ML can occur in many different types and forms. Certain scenarios, e.g., real estate fraud, are highly specific with only a few transactions occurring. However, banks are usually more involved in such processes than a bank solely executing the transaction, making them more informed and capable of assessing ML risks than if they had only executed the transaction. Such transactions must be manually assessed by experts and therefore TM systems are of limited use. To further elaborate on this, we now turn to an in-depth discussion of ML typologies specifically related to bank transactions.

The concept of typologies concerning ML defines how money can be laundered using bank transactions. Some typologies have already been around for quite some time, while new ones have occurred with, for example, cryptocurrencies becoming more mainstream (Elliptic, 2020). Table 1 below provides an overview of known transactional ML typologies. A subset of the three most important typologies, which have been determined by insights gained from the interviews, will be evaluated during the course of this study.

Table 1. An overview of bank transaction money laundering typologies.

Typology	Description	Reference
Low to high	Accounts that initially register low transaction amounts that soon increase to substantial amounts.	UNODC[1], Wheeler
Structuring/ smurfing	Dividing large sums over multiple transactions to remain below reporting thresholds. This can be combined with almost all other typologies. With *smurfing*, usually structuring large sums using multiple individuals is meant.	UNODC, FIAU[2], bronID[3], EY[4], BIS[5], Wheeler[6], AMLC[7]
Geographical structuring/fan-out	Spread transactions through multiple offices, countries, or people to attract less attention or to make tracing harder. Combined with fan-in as second step this is called Scatter-Gather.	UNODC, FIAU, bronID, EY, IBM[8], Sun et al.[9]
Bipartite	Multiple originating accounts divide flows over multiple receiving accounts, potentially combined with *Unrelated people.*	IBM
Type change	Change in type of usual transaction (e.g., cash or transfer) to receive or transfer money.	UNODC, EY
Cash	Unusual high deposits or withdrawals immediately or over a brief period. Usually combined with another typology.	UNODC, bronID, EY, BIS, Sun et al., Wheeler
Inactive/dormant	Accounts that have long been inactive and suddenly receive deposits, potentially combined with *Cash.*	UNODC, BIS
Wallet	Accounts that only register deposits over a certain period, potentially combined with *Cash*.	UNODC, Sun et al.
International trade	Accounts receiving from or transferring to countries other than the one importing/exporting goods or providing services.	UNODC, BIS
Money mule/fan-in	Accounts in the name of multiple people/organizations with the same person or party having actual control. Combined with fan-out as second step this is called Gather-Scatter.	UNODC, bronID, BIS, IBM, Sun et al.
Pass through/ funneling	Accounts registering debit and credit transactions of the same or similar amounts, indicating a cover-up account.	UNODC, EY, Wheeler
Profile inconsistency	Individuals or organizations having higher cashflows through accounts than financial or commercial statements can explain.	UNODC, FIAU, bronID, EY, Wheeler
Changing information	People (frequently) changing personal information at the time of transferring.	UNODC, EY, Wheeler

[1] (UNODC, 2010)
[2] (FIAU, 2021)
[3] (BroniD, 2019)
[4] (EY, 2018)
[5] (Bank for International Settlements, 2023)
[6] (Wheeler, 2021)
[7] (AMLC, 2020)
[8] (Suzumura & Kanezashi, 2018/2023)
[9] (Sun et al., 2022)

Unrelated people	Transfer receiving from someone random who is not related without reasonable justification.	UNODC, bronID
Currency exchange	Large and continuous purchase transactions for foreign currency or cryptocurrency.	UNODC, bronID, EY
High-Risk Jurisdictions	Transfers to or from (parties in) tax havens or other high-risk countries.	FIAU, bronID, Wheeler
Circular/cycle	Round-tripping transactions where funds end in the originating jurisdiction after being transferred to a foreign country, potentially combined with *High-Risk Jurisdictions.*	FIAU, bronID, EY, IBM, Wheeler
Laundromat	Front companies default on fake loans, then authenticated to debt by corrupt judges allowing payment from 3rd party, potentially combined with *High-Risk Jurisdictions*.	Harding[10]
Trade-based ML (e.g., Black Market Peso Exchange)	Trade is under-, over- or double invoiced, creating illicit money flows between organizations or jurisdictions.	FinCEN[11], FATF[12], Cassara[13]

2.4 Indicators of Money Laundering

This subsection presents the indicators as set out in law and described in literature which can be used by Dutch banks to determine whether a transaction is unusual and must be reported. To unpack this topic further, let's differentiate between objective and subjective indicators.

2.4.1 Objective and Subjective Indicators of Unusual Transactions

Dutch banks are legally required to report unusual transactions to the Dutch FIU (Wet ter voorkoming van witwassen en financieren van terrorisme, n.d.). If the FIU declares a reported transaction suspicious, banks receive a so-called dissemination notice that only states that the transaction has been declared suspicious, but not why (FIU-Nederland, 2021). However, while this notice itself contains limited information, the fact that a reported transaction has been declared suspicious already provides useful information on its own. The same applies to information requests originating from criminal investigations by authorities.

To determine whether a transaction is unusual or not, Dutch banks can use six indicators established by the legislature in the 'Wwft Implementing Decree' (Uitvoeringsbesluit Wwft 2018, n.d.). Table 2 below presents those indicators.

Table 2. Unusual transaction indicators for Dutch banks

Indicator	Description
Subjective01	A transaction that the institution has reason to believe may be related to ML or terrorist financing.
Objective01	It stands to reason that transactions reported to the police or Public Prosecution Service in connection with ML or terrorist financing should also be reported to the Financial Intelligence Unit; after all, there is a presumption that these transactions may be related to ML or terrorist financing.
Objective04	A transaction for an amount of €10,000 or more, involving cash exchange into another currency or from small to large denominations.

[10] (Harding, 2017)
[11] (FinCEN, 2010)
[12] (FATF, 2006)
[13] (Cassara, 2015)

Objective05	A cash deposit for an amount of €10,000 or more in favor of a credit card or a prepaid payment instrument (prepaid card).
Objective06	The use of a credit card or prepaid payment instrument (prepaid card) in connection with a transaction for an amount of €15,000 or more.
Objective12	A transfer of money for an amount of €2,000 or more, unless it concerns a transfer of money by an institution that has entrusted the settlement of the transfer of money to another institution that is also subject to the reporting obligation, as referred to in Article 16, paragraph 1, of the Act.

As Table 2 above shows, all except one indicator are of an objective nature. This means that Dutch banks are required to report transactions that meet one or more of those objective indicators either way, regardless of the circumstances, or without having to investigate them for 'unusualness'. This implies that banks should very easily be able to avoid false negatives for those transactions since the objective indicators can easily be tested for. Therefore, the focus of this study is primarily on transactions that should have been reported based on the subjective indicator.

The reports based on objective indicators are mainly used by the FIU to gain general intelligence but are often less indicative of actual ML occurring than reports based on subjective indicators as they have already been declared unusual by a bank (FIU-Nederland, 2022). The one subjective (Subjective01) indicator, however, is much harder to assess, as a bank has to determine for themselves whether they consider a particular transaction unusual. Expanding on this point, the following subsection presents transactional indicators banks can use within their rule-based transaction monitoring systems.

2.4.2 Operational Indicators of Money Laundering Through Bank Transactions

The primary indication of ML concerning bank transactions occurring is 'unusual' activity in an account. Having outlined the objective and subjective indicator(s), we now segue into a review of the existing literature on operational indicators within banks. Although distinguishing this uncommon behavior is not an easy task, certain attributes of accounts and transactions can indicate unusual transactions associated with ML. An analysis of case histories as disclosed by the Dutch FIU and reports as published by other authoritative bodies presents the following list of indicators relevant to the Dutch financial system, outlined in Table 3 below.

Table 3. Indicators of transactional money laundering in the Netherlands

Indicator	Description	Reference
Type	Type of transaction. Can be a 'normal' bank transfer, cash deposit or withdrawal or transfer via another payment provider.	FIU[14], EBA[15], AMLC[16]
Amount	The amount of the transaction in relation to the static (e.g., student account) or dynamic threshold.	FIU, EBA, AMLC, FATF[17], DNB[18]
Frequency	The frequency of the transaction (potentially of a similar amount) in relation to the static or dynamic frequency threshold.	FIU, EBA, FATF, FINTRAC
Sector salaries	Above average salary payments for a certain sector.	FIU

[14] (FIU-Nederland, n.d.-a)
[15] (European Banking Authority, 2018)
[16] (AMLC, 2020)
[17] (FATF – Egmont Group, 2020)
[18] (DNB, 2020)

Sector high-risk	Organizations active in high-risk sectors vulnerable to ML. Or payments to or from high-risk sectors.	FIU, EBA, FATF, DNB, FINTRAC
Volatility	Relatively large changes in the account balance over a brief period.	FIU, EBA, FINTRAC
International exposure	Percentage of in or outflow in relation to static or dynamic thresholds.	FIU, AMLC, FATF, FINTRAC
International high-risk	Organizations active in high-risk countries vulnerable to ML. Or payments to or from high-risk countries.	FIU, EBA, AMLC, FATF, DNB, KLPD[19]
Account type	Personal (e.g., student or normal) or business account.	FIU, FINTRAC
Account duration	Extremely high account balances on recently opened accounts.	FIU, EBA, FATF, DNB

2.5 Related Work

Extensive research has been undertaken to enhance AML TM systems through the application of both supervised and unsupervised machine learning classifiers. Key contributions to this domain can be categorized into two primary advancements: graph-based machine learning and various other deep learning approaches. The former involves representing bank clients and their accounts as nodes within a graph to facilitate network structure analysis. This approach used for AML purposes was first described in foundational research by Weber et al. (2018). These authors also developed the AMLsim simulator, a multi-agent platform for generating synthetic financial datasets for AML research. AMLsim enables the creation of large-scale, realistic financial networks, simulating the real-world behavior of entities engaged in typical and ML transactions. It creates a detailed graph of financial interactions, with nodes representing bank accounts and edges symbolizing transactions. This setup allows for the exploration of typical and suspicious transaction patterns in a controlled, large-scale environment.

Subsequently, Eddin et al. (2022) proposed a model to optimize the risk assessment of generated alerts using this graph-based method. On the other hand, the latter approach harnesses other deep learning methods, most notably neural networks, as evidenced by works such as Han et al. (2020) and Zhang & Trubey (2019). Rocha-Salazar et al. (2021) utilized this approach to develop a clustering model, aimed at enhancing both self-comparative and group-comparative analyses of clients to identify potentially suspicious ML transactions. While scholarly attention often leans toward minimizing false alerts, the primary emphasis tends to be on overall model accuracy, targeting a reduction of both false positives and false negatives (Jullum et al., 2020). Notably, model recall, or the focus on reducing false negatives, is less emphasized in the development of novel models and approaches, even when working with highly imbalanced datasets. To the best of our knowledge, no existing study has specifically investigated the feasibility of estimating the FNR in rule-based TM systems within the AML context, marking a clear gap in the academic literature. Therefore, our work diverges from this focus by explicitly aiming to quantify the FNR of rule-based TM systems.

As the paragraph above indicates, current research within this domain has primarily been focused on further enhancing and refining machine learning detection algorithms for TM purposes. Conversely, despite the extensive use of rule-based TM systems in current practice, there is a lack of comprehensive research on the evaluation of those systems. The academic community has not adequately explored how well these rule-based systems perform, especially in terms of transactions that they incorrectly fail to flag for review. This lack of interest persists whether the objective is to enhance these systems directly or to derive a more generalized

[19] (KLPD - Dienst Nationale Recherche Informatie, n.d.)

understanding of their workings and limitations. This oversight represents a significant area for further investigation, as knowing the FNR in the context of existing rule-based systems is crucial for enhancing overall TM effectiveness in the banking sector. As such, there is a need to improve the evaluations of rates of rule-based TM models regarding transactions that have inaccurately been left unreviewed, a critical metric that has largely been overlooked up to now.

Although the feasibility of estimating the FNR in rule-based TM systems within the AML context has not yet been studied, significant work has been done in other fields that grapple with false negative classification errors and false negative estimation in large, skewed-class datasets. This has primarily been the case in the field of ecological research focusing on animal populations, where comprehensive individual examination is often impracticable or even impossible. To address this issue, Petersen founded the capture-recapture method as early as 1896, initially aiming to estimate fish population sizes (Southwood & Henderson, 2009). This method employs two independent classifiers to estimate various parameters, including false negative predictions, of models for largely unobserved populations. Over time, the applicability of the capture-recapture method for estimating false negatives and false negative rates has expanded, finding utility in diverse areas such as medical screening (Goldberg & Wittes, 1978), road traffic safety (Abegaz et al., 2014; Razzak & Luby, 1998) and social sciences (Brittain & Böhning, 2009). Building upon this, Mane et al. (2004) extended the method to create a systematic approach for estimating false negatives in generalized two-class classification problems, exemplifying its efficacy on a highly imbalanced dataset of spam emails.

Additionally, the estimation of the FNR has also been researched using other methods. Connors et al. (2014) studied the frequency and magnitude and provided quantitative assessments of both false-positive and false-negative observation errors in the context of classifying the threat and recovery status of animal populations in nature. They estimated the FNR by comparing the predictions from both a simulation model with the actual underlying trend as well as a model devoid of such a trend. The authors quantified the FNR as "the proportion of simulations where we failed to reject the null hypothesis …, when in fact the null hypothesis was false.". A unifying thread across literature employing these methods is the shared objective of estimating false negatives and FNRs using classifiers for predominantly unobserved populations—a challenge that aligns directly with the objective of this study focused on potential ML transactions in high volume datasets.

In summary, existing literature provides various methodologies and models suited to estimate the FNR of rule-based TM systems in ML classification. Despite this, as current academic endeavors predominantly concentrate on enhancing TM systems through the innovation of novel machine learning ML detection algorithms, there exists a noticeable gap in the literature that directly addresses rule-based TM systems which are now primarily used within Dutch banks. Therefore, this study serves to bridge this significant gap in the academic landscape by empirically exploring the feasibility of estimating the FNR of rule-based TM systems in the context of high-volume ML transaction classification using machine learning classifiers.

2.6 Statistical Methods

In order to evaluate the suitability of historical alerts for estimating the overall FNR pertaining to unreviewed transactions, this study employs supervised machine learning classifiers designed to discern the correlation between the characteristics of historic alerts and the probability of actual ML involvement in those transactions. Afterward, these trained models can be used to predict whether unreviewed transactions constitute ML.

2.6.1 Machine Learning Classifiers

This subsection presents a short description for every supervised machine learning classifier used in this study to train and predict whether unreviewed transactions are indicative of ML. Since the goal of this study is not to develop new machine learning (potentially deep learning) ML detection algorithms which beat current benchmarks, we want our classifiers to be relatively

simple and computationally cheap to work with. Especially considering the size of the datasets used throughout this study and in practice.

The most conventional method to predict the probability of a binary event is Maximum Likelihood Logistic Regression (Zhang & Trubey, 2019). Using a parametric statistical model, it predicts the probability of a dependent output class (ML or not) using one or multiple independent input variables. The output is a probability on the interval between 0 and 1 of an event (such as "indicative of ML) occurring.

The second classifier are decision trees, as per Alkhalili et al. (2021). These trees are constructed by employing a splitting technique that determines the input attributes and progressively moves through the training data to achieve the desired output. Within the structure of a decision tree, two fundamental entities emerge: decision nodes and leaves. The decision nodes serve as points of data division, guiding the tree's branching based on specific criteria, while the leaves represent the ultimate decisions or outcomes associated with the given data configurations. The decision tree algorithm's ability to handle both categorical and numerical data without requiring preliminary transformation is advantageous, reducing data preprocessing needs.

Third, ensembles of decision trees, as highlighted by Bhattacharyya et al. (2011) in general exhibit better performance compared to other techniques, capturing fraudulent cases effectively while minimizing false positives. To this end, we implement models of the following types: 'standard' random forest, (balanced) bagged decision trees, balanced random forest, the AdaBoost implementation using decision trees and RUSBoost. Furthermore, ensembles of decision trees inherently handle imbalanced data well, which is a common challenge in ML detection, thus making it a robust choice for estimating the FNR of TM systems in Dutch banks. Additionally, the computational efficiency and simplicity of implementation of random forests make them highly practical for deployment on large datasets. In addition, we have also used balancing measures throughout the models, as elaborated on in 3.2.2 Machine Learning Classifiers.

2.7 Chapter Conclusions

- The predominant focus in existing AML research centers on developing novel machine learning classifiers.
- A notable research gap exists in the evaluation of rule-based TM systems, although they are extensively used in practice. This gap is particularly evident in the exploration of the FNR of these systems, an aspect often overlooked in current research.
- Methods from other disciplines like ecology offer potential for estimating the FNR in imbalanced ML transaction datasets.
- We employ Logistic Regression, Decision Trees and various ensembles of Decision Trees as classifiers to estimate the FNR pertaining to unreviewed transactions.
- We use synthetic datasets with ground truth labels to enable FNR evaluation and performance comparison, which we evaluate primarily using the AUPRC metric Matthews Correlation Coefficient.

3. Methodology

This chapter presents the research approach and research methods utilized for conducting the study. For the latter, it therefore also provides information on the statistical metrics and synthetic data used.

3.1 Research Approach

To explore the theoretical feasibility of estimating the FNR pertaining to unreviewed transactions, this study adopted both qualitative and quantitative methods. This combination in a concurrent mixed methods research approach, as endorsed by Creswell & Creswell (2017, p. 14), offered the advantage of enabling the integration of specialist knowledge from the field with generalizable results from statistical tests. This methodology enabled better guidance from specialists which enhanced the accuracy of modeling rule-based TM systems. Additionally, this approach supported the triangulation of knowledge from specialists and quantitative statistical results, thereby augmenting the study's comprehensiveness and validity (Jick, 1979; Morse, 1991).

First, qualitative data were collected from literature and interviews with specialists. These insights provided guidance for modeling the behavior of rule-based TM systems with greater accuracy. Informed by these qualitative insights, the study shifted its focus to quantitative modeling and statistical analysis. To enable empirical calculation of the actual FNR of rule-based TM models, the use of synthetic data was deemed indispensable, especially given that the existing knowledge gap stems from the largely unknown performance of these models on real-world data. Consequently, with the aim of determining the theoretical feasibility of estimating the FNR pertaining to unreviewed transactions, multiple types of statistical models were developed. Subsequently, the FNR of their predictions was compared against the pre-determined actual FNR of the rule-based TM system, which was known in advance. In doing so, the study ascertained the potential for historical transaction alert data to offer a reliable estimate of the FNR related to unreviewed transactions.

3.2 Research Methods

Following the mixed method research approach, both quantitative and qualitative research methods were applied. Sub-questions 1, 2 and 3 involved qualitative methods such as a literature review and semi-structured interviews. This enabled an in-depth understanding of the current systems and their limitations but did not provide all the information necessary to answer the main research question. Therefore, to address sub-question 4, the study employed synthetic transactional datasets for simulating alerts and utilized supervised machine learning classifiers to predict for unreviewed transactions whether they were indicative of ML to enable FNR calculations.

3.2.1 Interviews

Seven interviews were conducted with specialists in the field, all following a semi-structured approach. This approach facilitated the collection of additional contextual information alongside responses to predetermined interview questions (Adams, 2015). The goal of these interviews was to gather insights on transactional ML typologies and the workings of rule-based TM systems. The set of interview questions employed is presented in Appendix B. Interview Questions. Of the seven interviewees, six have hands-on experience with rule-based TM systems in banks, while the other participant has experience working at a Dutch regulatory body. All specialists with expertise in rule-based TM systems at the organization of the internship position were interviewed. In addition to these, two specialists at other organizations, who were contacted via the organization of the internship position, were consulted to gain a broader perspective. Among the interviewees, four had either current or past employment at a bank focusing on rule-based TM systems. All interviews, except for one, were conducted in person. Furthermore, except for

two cases constrained by the limitations of the interview locations, physical interviews were recorded for subsequent playback and analysis.

The interviews primarily revolved around key themes such as common transactional ML typologies in Dutch banking, indicators for recognizing these typologies and the implementation and effectiveness of threshold-based rules in TM systems. The interviewees provided insights on the nature of static and dynamic rules, the complexity of rule parameters and the variance in risk assessment methodologies employed by banks.

3.2.2 Machine Learning Classifiers

This methodology section outlines the key characteristics and configurations of each classifier, aiming to detect ML activities within bank transactions. We used eight different base classifier types, which selection was driven by their suitability for handling imbalanced datasets. To address the challenge posed by imbalanced datasets, we also implemented balancing measures in certain classifiers that offer this functionality to gain better performance. For these classifiers, we implemented them with both their default settings to create a baseline performance as well as with the custom balancing measures applied to determine the performance gains. The measure available for most models was to implement cost-sensitive learning. This penalizes misclassifications of the minority class much higher, therefore prioritizing the model towards correct classifications and detection of this class. This first measure was available for and applied to five out of the eight model types used in this study. The second measure available was to first downsample the majority class in the train dataset, in order to mitigate the class imbalance. This second measure was available for four out of the eight model types. The choice of these measures is driven by their potential to enhance model performance in scenarios typical of TM systems used by Dutch banks. Table 4, presented below, provides a comparative overview of all seventeen model configurations along with their balancing measures combinations. The balancing features are further discussed in detail in 5.2.5 Experimental Setup. We used a set random state throughout all model runs and downsampling efforts.

Table 4. All classifier model setup combinations

Model	Default	Cost-sensitive learning	Downsampling
Logistic Regression	x	-	-
Logistic Regression balanced	-	x	-
Decision Tree	x	-	-
Decision Tree balanced	-	x	-
Decision Tree bagged 10 estimators	x	-	-
Decision Tree bagged balanced 10 estimators	x	-	x
Decision Tree bagged 100 estimators	x	-	-
Decision Tree bagged balanced 100 estimators	x	-	x
Random Forest	x	-	-
Random Forest balanced	-	x	-
Random Forest balanced subsample	-	x	-
Balanced Random Forest	x	-	x
Balanced Random Forest balanced	-	x	x
Balanced Random Forest balanced subsample	-	x	x
AdaBoost	x	-	-
AdaBoost with Decision Tree balanced	-	x	-
RUSBoost	x	-	x

Following is a detailed exploration of each base classifier type utilized in our study. This section will delve into the specific configurations and applications of these classifiers, highlighting their roles in addressing the challenge of detecting ML activities within bank transactions

Logistic Regression

Logistic Regression is a statistical model used for binary classification (Scikit-learn developers, n.d.-f). It predicts the probability of a binary outcome based on one or more independent variables. In this study, it is configured with a maximum of 2000 iterations, which has been proven sufficient to let the model converge for all experiments. Additionally, a balanced version using cost-sensitive learning is also implemented.

Decision Tree

The Decision Tree Classifier is a simple and effective non-parametric supervised learning method used for classification (Scikit-learn developers, n.d.-e). It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A balanced version using cost-sensitive learning is also implemented.

Bagging with Decision Trees

To enhance the stability and accuracy of decision trees, the ensemble method of bagging (Bootstrap Aggregating) is employed (Scikit-learn developers, n.d.-c). This combines the predictions from multiple decision trees. This study uses two variants: one with 10 estimators and another with 100, aiming to capture diverse decision boundaries and reduce overfitting. A balanced version using a downsampling strategy is also implemented.

Balanced Bagging with Decision Trees

Similar to Bagging with Decision Trees, this model combinations the predictions from multiple decision trees (Imbalanced-learn developers, n.d.-a). However, this type also incorporates an additional downsampling balancing step on every training set.

Random Forest

This classifier is known for its robustness and ability to handle large datasets with numerous input variables (Scikit-learn developers, n.d.-g). Its ensemble approach is a specialized form of bagging with Decision Trees, aggregating multiple decision trees too. However, it introduces randomness in the selection of features to increase the generalizability. Additionally, a balanced version using cost-sensitive learning is also implemented.

Balanced Random Forest

To address the challenge of class imbalance, the Balanced Random Forest Classifier is deployed (Imbalanced-learn developers, n.d.-b). The default balanced Random Forest uses downsampling strategies to mitigate class imbalance. A variant which adjusts class weights as well is also implemented.

AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble method that combines multiple weak learners to create a stronger model (Scikit-learn developers, n.d.-b). It is used for its ability to sequentially correct the mistakes of its components. Two versions are used: one is the standard AdaBoost Classifier, and the other employs a Decision Tree classifier with balanced class weights as the base estimator.

RUSBoost

Integrating Random Under-Sampling (RUS) with the AdaBoost algorithm, RUSBoost is particularly adept at handling class imbalances (Imbalanced-learn developers, n.d.-d). Its inclusion in this study is driven by its capability to focus on the minority class.

This overview of the various classifiers and their configurations underscores the multifaceted approach we have adopted to tackle the complexities of TM. Next, we will transition to an in-depth discussion of the synthetic datasets used.

3.3 Synthetic Data

This section explores the creation, characteristics, and role of the synthetic datasets in the research process. To be able to create a statistical model based on simulated alerts, a transactional dataset that mirrors the properties of actual bank transactions in the Netherlands was needed. However, due to privacy concerns, it was virtually impossible to obtain real-world transactional datasets from financial institutions. In addition, as the accuracy of ground-truth labeled data within real-world datasets concerning ML is uncertain, a synthetic dataset where those labels are perfectly known in advance is better suited for testing the theoretical feasibility of the concept.

Several synthetic transactional datasets and simulators that can incorporate ML behavior were available for this purpose. Transactions within those datasets are dichotomous labelled as ML or not. This label will be assumed to correspond with the 'unusual' label in real-world historically flagged transaction alerts. Examples of already existing open-source data simulators are the PaySim simulator from Lopez-Rojas & Axelsson (2017/2023; 2016) and the AMLsim simulator from IBM (Suzumura & Kanezashi, 2018/2023). Additionally, published pre-generated datasets from these and other simulators are publicly accessible. Notable examples include those generated by J.P. Morgan (n.d.) and a more advanced version of IBM's AMLsim simulator called IT-AML (Altman, 2023; Altman et al., 2023).

After a thorough comparison, as shown in Table 4Table 5 below, the pre-existing published datasets generated using the IT-AML simulator were deemed most suitable to conduct further modelling on. The PaySim simulator proved unfit for this research as it was designed solely for the simulation of mobile money transfers. In addition, it focuses on generating general fraudulent transactions rather than those specifically indicative of ML. The transactional dataset published by J.P. Morgan was deemed unsuitable as well because it lacks detail concerning the typologies used for embedding ML behavior in transactions and incorporates only a single currency. Given that the study aims to assess a select subset of ML typologies, the absence of this crucial information disqualified the dataset. While the AMLsim simulator does include information on the ML typologies used in transaction generation, it presents limitations. Notably, half of these typologies are labeled as 'currently under construction,' with no further clarifications offered. The simulator further supports only a single transaction type and currency. Moreover, the open-source version of AMLsim has languished in terms of maintenance and updates due to its evolution into the closed-development IT-AML simulator for a considerable duration.

In comparison with the simulators and datasets elaborated on above, the pre-existing published datasets generated via the IT-AML simulator have proven to be most suited for this research since they contain multiple currencies and types of transactions and provide information on ML typologies used. In addition, they are available in three sizes with both a low and high ML prevalence rate and incorporate the most advanced ML behavior from available simulators and datasets. Therefore, we selected those datasets to conduct further modelling on.

Table 5. Overview of comparison between various synthetic datasets and simulators

Dataset	Suitability for ML Modelling	Diversity in Transaction Types & Currencies	Detail in ML Typologies	Maintenance & Updates
IT-AML Simulator	High suitability	Adequate diversity	High detail	Well-maintained and updated
PaySim Simulator	Low suitability (focused on mobile money transfers)	Limited (mobile money transfers only)	Low detail (general fraudulent transactions)	Not specified

J.P. Morgan Dataset	Low suitability (lacks detail in ML typologies)	Limited (single currency)	Low detail	Not specified
AMLsim	Moderate suitability	Adequate diversity	Moderate detail (half typologies under construction)	Lacking (evolved into closed-development IT-AML)

3.3.1 IT-AML Simulator Datasets

The IT-AML datasets, as published by IBM in February 2023, were generated using their most advanced transactional simulator capable of incorporating ML behavior and are publicly accessible. The datasets serve as a simulated representation of a complex financial ecosystem, incorporating various types of accounts - namely individuals, corporate entities, and financial institutions. The interactions among them extend from individual-to-individual financial exchanges to more complex dealings between individuals and enterprises, as well as intra-enterprise transactions. These transactions manifest in diverse forms and include activities such as the procurement of consumer goods and services, the issuance of industrial supply purchase orders, the disbursement of salaries and the fulfillment of loan repayment obligations. These financial activities are intermediated through banks, wherein both the transaction initiators and recipients maintain a variety of account types. These range from traditional checking and credit card accounts to more contemporary financial instruments like cryptocurrency wallets. As a result, the datasets present a comprehensive snapshot of a diverse financial ecosystem.

A small proportion of the accounts, both personal and business, within the datasets participate in unlawful activities. The illicit funds, obtained from these illegal activities are then concealed via a series of financial transactions. Every transaction involved in this series therefore constitutes ML and is accordingly tracked and labeled within files supplied with the datasets. The generator simulates all three phases of the ML cycle: placement (source of illicit funds), layering (mixing them into the financial system) and integration (spending the funds).

One notable advantage of utilizing synthetic data is the ability to overcome limitations inherent to real transactional data. In the context of banking institutions, it is important to note that these organizations, and therefore their models, typically have access to only a subset of transactions related ML, namely those involving their own institution. Transactions happening at other banks or between other banks are not seen. Consequently, models constructed solely based on transactions from a single institution possess a restricted perspective of the overall financial landscape. In contrast, synthetic transactions in the IT-AML datasets encompass an entire financial ecosystem, providing a comprehensive representation. This comprehensive nature of synthetic data enables exploring the potential benefits of TM systems that possess a holistic understanding of transactions across multiple institutions. This is especially important as Dutch banks are increasingly collaborating to combine transaction data in one system, called 'Transactie Monitoring Nederland' (Transactie Monitoring Nederland, n.d.). This synthetic data therefore enables us to already provide insights into the benefits of this collaboration. We will test for both the perspective of the most present bank in the data as well as all the banks combined.

IBM has released the IT-AML dataset in the three sizes small, medium and large with a timespan of respectively 10, 16 and 97 days. Each size is available in a version with both a low and high ML prevalence rate. As an extended timespan in the dataset is beneficial for capturing a wider variety of transaction behaviors, which is essential for training machine learning classifiers to recognize subtle and complex ML patterns, datasets with a longer timespan were preferred. In addition, selecting same size datasets with different rates opens up possibilities for a direct comparison between classifiers when performing on those different prevalence rates. Therefore, the two large datasets with both the low and high ML prevalence rate have been selected.

Both datasets incorporate the same eight ML patterns, as presented in Figure 1 below (Altman et al., 2023). The patterns present transaction behaviors typically observed in ML scenarios. The Fan-Out pattern is characterized by a single source account dispersing funds to multiple destinations, while its counterpart, the Fan-In pattern, shows the consolidation of funds from various sources into one account. The Gather-Scatter pattern combines these two, illustrating funds being first accumulated and then distributed, whereas the Scatter-Gather pattern reverses this flow, depicting funds being dispersed first and then gathered. The Cycle pattern represents a closed loop of transactions where money returns to the originating account after passing through various others, embodying typical laundering schemes. In contrast, the Random pattern reflects a more haphazard flow, akin to a random walk, where funds do not return to the original account, often passing through entities like shell companies. The Bipartite pattern simplifies this, involving direct transfers from multiple sources to multiple destinations. Lastly, the Stack pattern adds complexity to this transfer model by introducing an additional layer of bipartite transfers, further obfuscating the fund's movement. These patterns collectively underline the sophisticated methods employed in ML, where entities control various accounts to manipulate transactions, challenging the enforcement and detection of illicit financial flows.

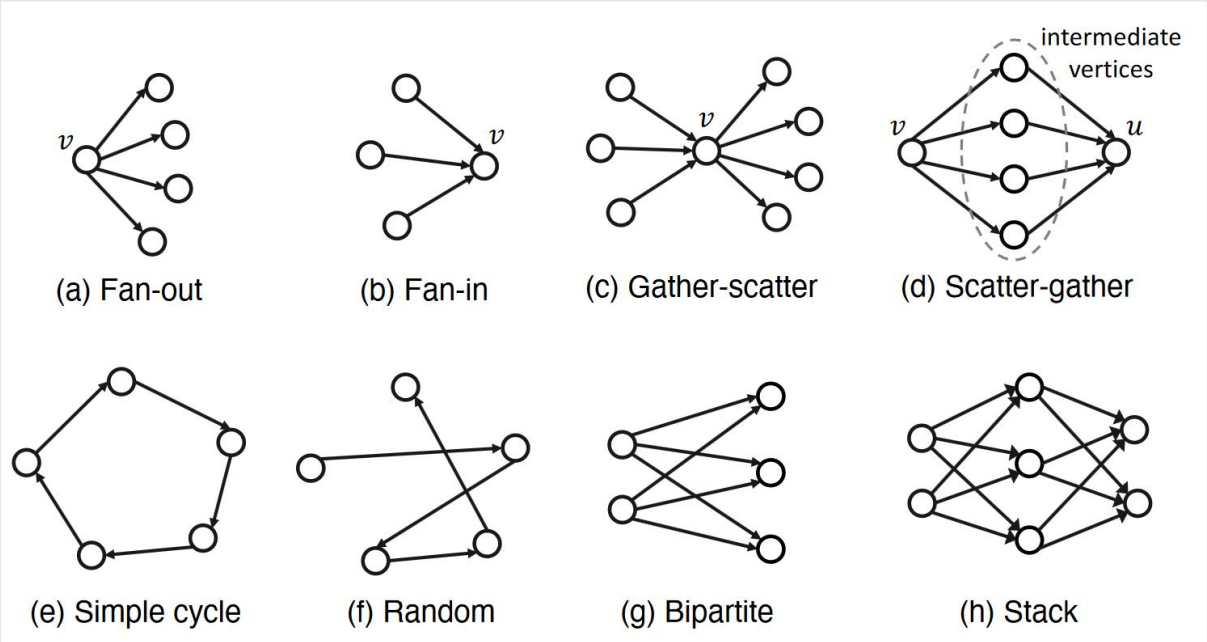

Figure 1. Available money laundering typologies within the dataset and simulator as reprinted from Altman et al. (2023)

Explorative Data Analysis and data preparation

The IT-AML datasets provide transactions in a tabular format, as outlined in Table 6. Next to the datasets itself, other adjacent files present which ML transactions belong to which typology. Metrics like the average transaction value, frequency, or the total value within a certain period can be computed. Behavioral nuances can be leveraged as well, such as the variance in transaction amounts or the number of unique transactional entities involved with an account, hinting at potential ML strategies like layering.

Table 6. IT-AML datasets format

Timestamp	From bank	Account	To bank	Account	Amount received	Receiving currency	Amount paid	Payment currency	Payment format	Is Laundering
2022/09/01 00:20	3208	8000F4580	1	8000F5340	0,01	US Dollar	0,01	US Dollar	Cheque	0
2022/09/01 00:26	12	8000EC280	2439	8017BF800	7,66	US Dollar	7,66	US Dollar	Credit Card	0

Table 7 presents an initial overview of the numerical characteristics of both the datasets.

Table 7. IT-AML datasets key overview

Measure	High prevalence rate	Low prevalence rate
Total number of transactions	179,702,229	176,066,557
Total number of ML transactions	225,546	100,604
Prevalence rate	0.13% (1/807)	0.06% (1/1,750)
Days spanned	97	97
Number of unique bank accounts	2,116,000	2,064,000

Since we are only interested in 'pure payment' transactions, all transactions of the 'Reinvestment' payment type have been excluded, removing 7,410,556 transactions from the high prevalence rate dataset and 7,258,238 transactions from the low prevalence rate dataset. The 'Bitcoin' payment, on the other hand, has been kept, as it will be considered a high-risk currency for the High-risk jurisdictions typology.

As Figure 2 shows, the large high prevalence rate dataset spans 97 days of normal behavior. The remaining days from 6th of November until the 12th of January only contain a very limited number of transactions, namely those in a ML series that has not yet been ended. Therefore, all transactions from the 6th of November have been excluded.

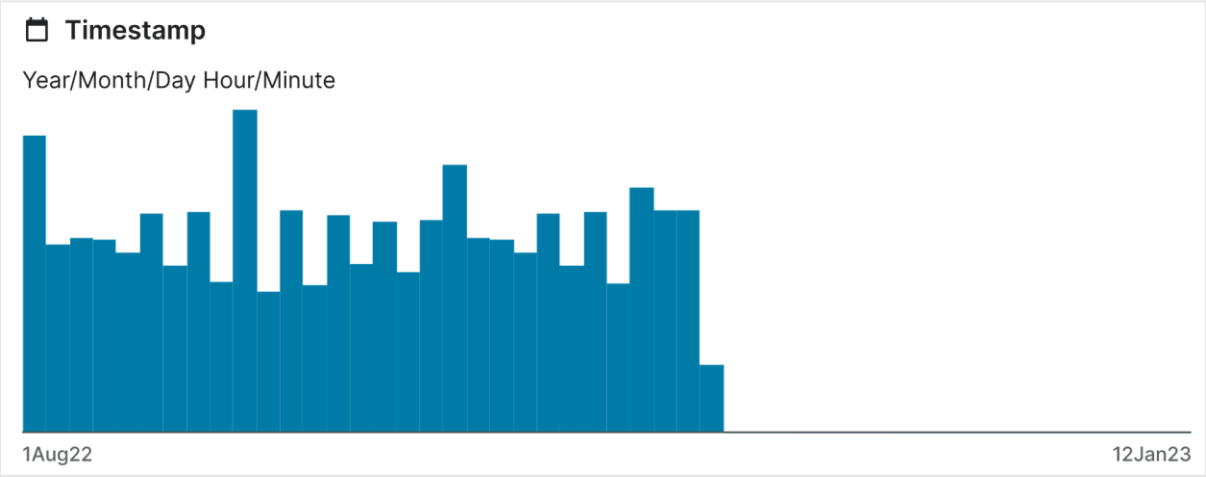

Figure 2. Timestamp visualization of the large IT-AML dataset as adapted from Altman (2023)

As we can see in Figure 3, the amount variables contain some very large outliers. They will be addressed, as discussed later in 5.1.1 Data Preprocessing.

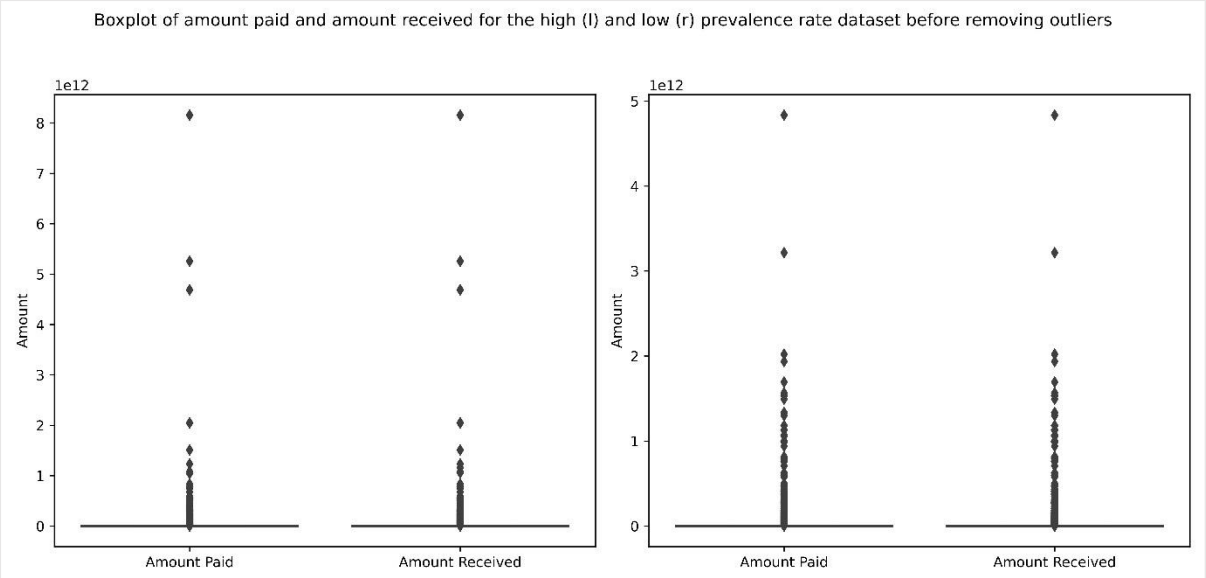

Figure 3. Boxplots of amount variables in both datasets before removing outliers.

The IT-AML datasets incorporate various type of payment formats, as depicted in Figure 4. This is particularly relevant because certain payment formats can be indicative of ML activities, or at least have a certain correlation with the likelihood of such activities. Notably, the presence of ACH transactions in these datasets indicates the inclusion of US-based financial activities. ACH, or Automated Clearing House, is an electronic network for financial transactions in the United States, commonly used for direct deposit, payroll, and vendor payments. Its presence in the dataset suggests that at least US transactions are represented, highlighting the international scope of the data.

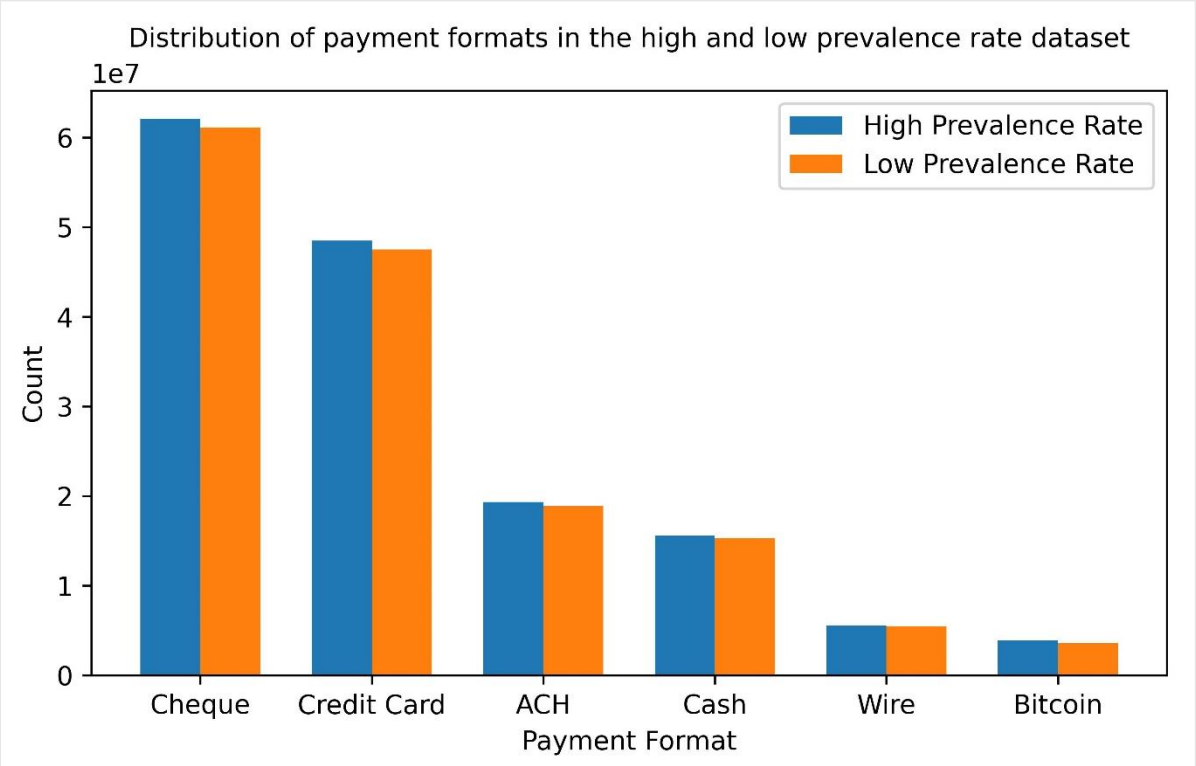

Figure 4. Distribution of payment formats across the two datasets.

The following Figure 5 strengthens this idea, as it clearly shows that the predominant currency for both receiving and payment transaction in both datasets is the US Dollar, with the runner up being the Euro followed by thirteen other currencies for a total of fifteen. We can use these currencies as a proxy for geographical data, which enables us to do flag transactions via countries based on their risk level or to flag cross-border transactions.

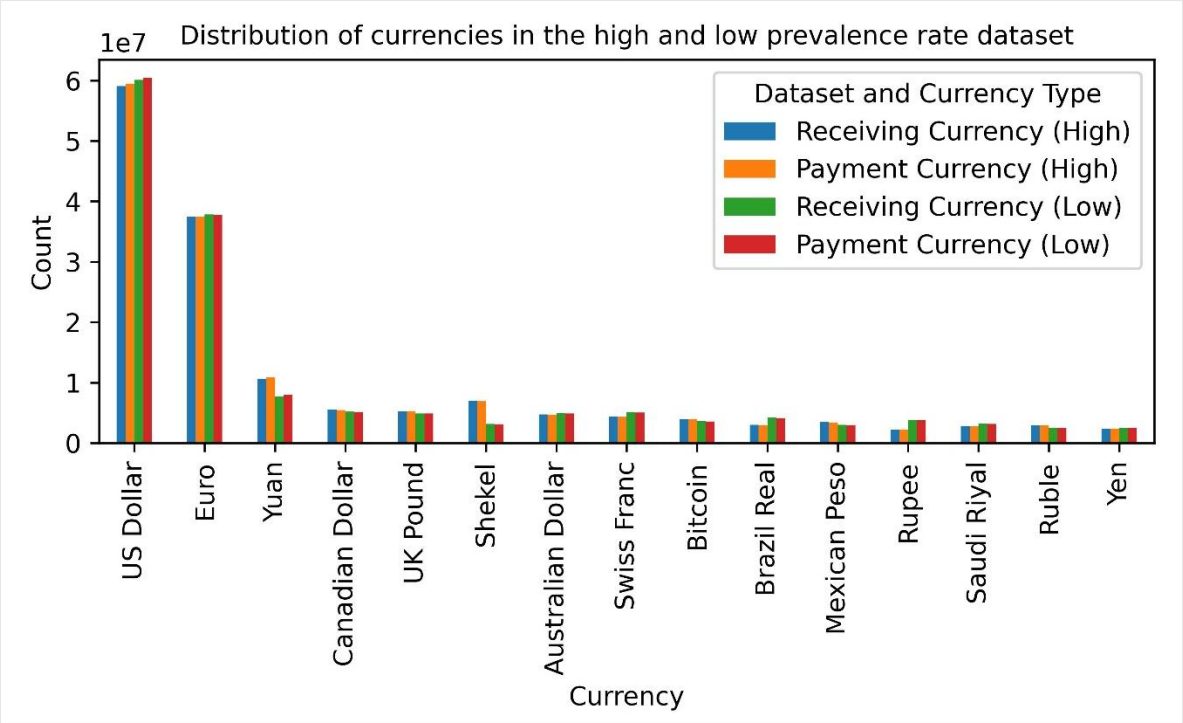

Figure 5. Distribution of currencies in both IT-AML datasets

The following Figure 6 and Figure 7 depict how often a bank was on the receiving or sending end of a transaction. We can clearly see that for the sending banks, there is one highly predominant bank present in both the datasets. This bank will therefore also be used for conducting experiments later on, as further explained in 5.2.4 Experiments. For the receiving banks, the distribution seems far more equally spread-out over all banks.

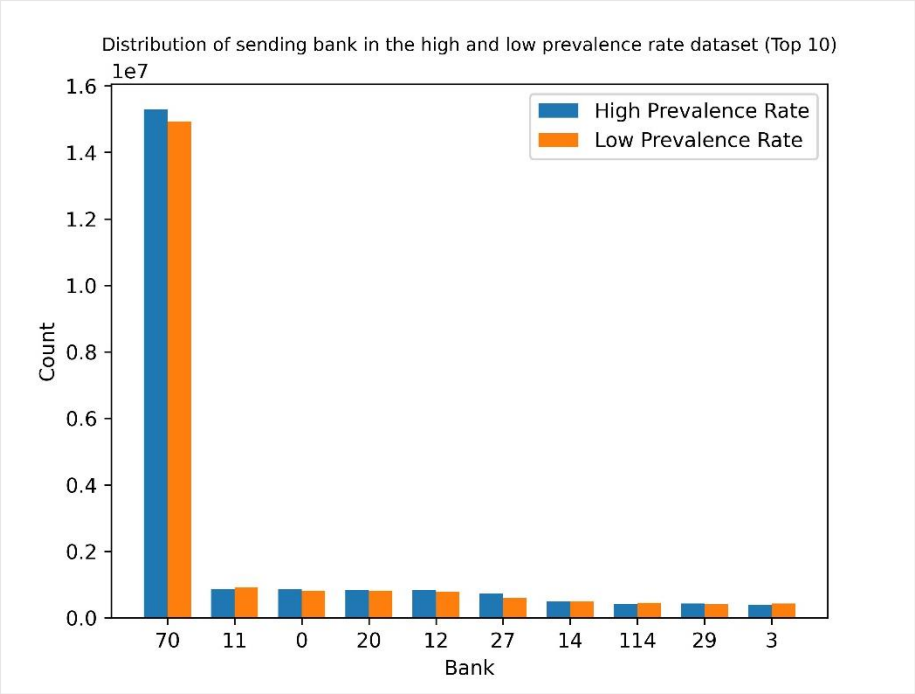

Figure 6. Top 10 banks where most transactions originate from.

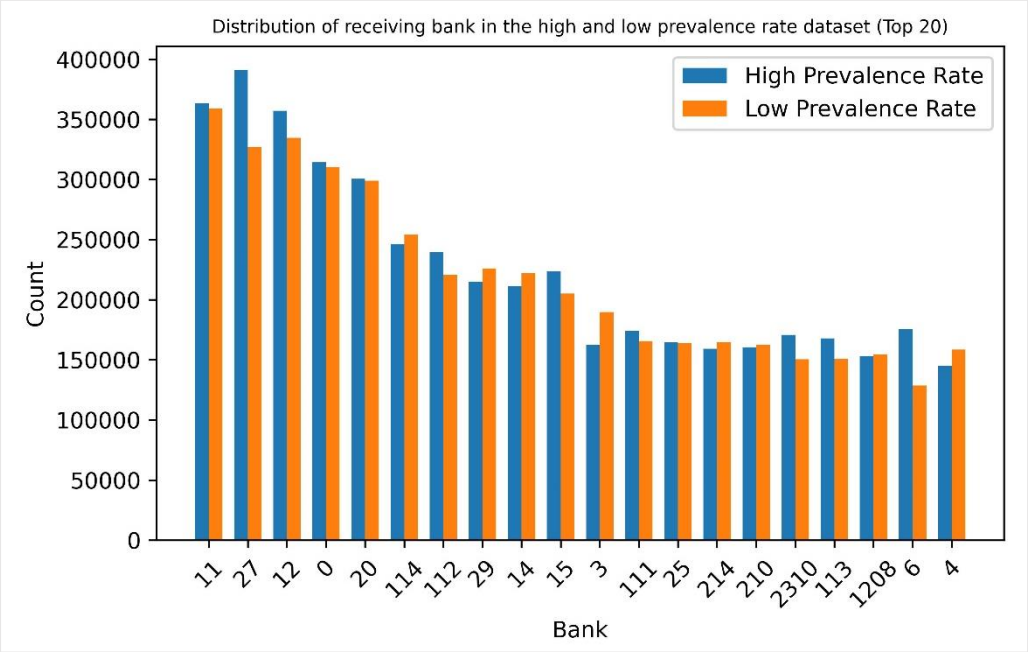

Figure 7. The top 20 banks on the receiving end of a transaction.

Both datasets incorporate the same eight ML patterns. Their count in both the datasets is presented in Table 8. Each pattern instance defines a series of ML transactions of that certain pattern.

Table 8. Count of money laundering patterns in the IT-AML datasets

ML patterns	High prevalence rate	Low prevalence rate
Bipartite	2109 (13%)	274 (12%)
Stack	2096 (13%)	259 (12%)
Cycle	2086 (13%)	298 (13%)
Scatter-gather	2067 (13%)	276 (12%)

Gather-scatter	2054 (12%)	284 (13%)
Fan-out	2040 (12%)	274 (12%)
Fan-in	2014 (12%)	278 (13%)
Random	2001 (12%)	275 (12%)

3.3.2 Tools

Model development and data analysis were conducted using the open-source coding language Python (version 3.11) to enable interoperability and ease of reproducibility (Python Software Foundation, 2023). Scikit-learn and Imbalanced-learn were the primary libraries for preprocessing data and training models (Imbalanced-learn developers, n.d.-c; Scikit-learn developers, n.d.-h). The software DataSpell has primarily been used for model development and data analysis (JetBrains, n.d.). For training the models and making predictions, a workstation with an Intel Xeon E7540 and 64Gb of RAM was used. We note that some datasets had to be sampled down due to calculations becoming too large for the available memory size.

4. Interview Insights

This chapter presents the insights obtained from the interviews conducted with seven experts working in the field of ML detection and prevention. To that end, it encompasses rule-based TM systems in general, the chosen subset of transactional ML typologies and the indicators that will be used to generate alerts.

4.1 Money Laundering in General

The first topic of exploration involves the complex and diverse characteristics of ML activities in Dutch society as highlighted by the interviewees. They noted that these features render comprehensive control exceedingly challenging (interviewee 4). Money launderers endeavor to simulate regular transactional behavior and illicit activities often involve smaller, less noticeable monetary amounts (interviewee 6). This is exemplified by the fact that an identical transaction could be indicative of ML for one client but entirely legitimate for another. Further complicating matters is that high-value transactions like mortgages or the sale of valuable assets are often legitimate activities (interviewee 6). These complexities are exacerbated by variations in risk associated with payments originating from different countries and sectors (interviewee 6). Consequently, banks face the formidable challenge of striking the right balance between finding "enough" true positives but also not generating too many false positives (interviewee 5).

Moreover, some ML activities are relatively simple to execute, further complicating regulatory efforts to mitigate them (interviewee 4). They do not only include (high value) criminal transactions but also a range of informal economic activities, such as employment in handyman or cleaning services, that go unreported for tax purposes, as well as commercial establishments that either underreport or entirely omit their cash revenue from official financial disclosures. Many petty criminal activities, such as the sale of stolen bicycles, predominantly involve cash transactions that often circumvent formal banking systems until eventually being used or deposited as ostensibly legitimate funds by other individuals or commercial entities like retail stores. It is important to underscore that relatively modest cash sums that remain undeposited generate no bank transactions and are consequently imperceptible to financial institutions as potential ML activities (interviewee 4). The same applies to alternative financial conduits such as underground banking systems, as well as transactions involving the exchange of high-value assets like art or automobiles as a form of payment. (interviewee 4)

Conversely, cash proceeds from more significant and severe types of crime are commonly sought to be laundered and deposited into bank accounts, as opposed to keeping them in cash, to facilitate their integration into mainstream financial activities (Interviewee 4). The Dutch government prioritizes the scrutiny of ML activities involving funds that originate from illicit criminal enterprises over those relatively small crimes, owing to the more substantial adverse impact such activities exert on society. For this reason, both regulatory authorities and banking institutions strategically concentrate their AML initiatives on transactions involving substantial sums that are associated with severe criminal conduct (interviewee 4). In summary, this first section has delineated the multifaceted nature of ML, making it clear why this area is challenging for financial institutions.

4.2 Rule-based Transaction Monitoring Systems

Building upon the challenges posed by ML, we delve into the mandated safeguards that financial institutions must have in place. Legislation requires these institutions to be ‘in control’ of ML activities as aligned with their risk appetite. Importantly, this does not automatically necessitate the implementation of TM systems (interviewee 5). Nevertheless, Dutch banks are obligated by the national regulatory authority, De Nederlandsche Bank (DNB), to formulate a Systematic Integrity Risk Analysis (SIRA). This comprehensive analysis is designed to identify various risks, including those related to ML, and prescribe corresponding mitigation measures (DNB, 2020). Notably, for larger retail banks, automated TM systems often serve as one such mitigation

strategy (interviewee 5). The SIRA is dependent on the markets in which a bank is active, the complexity of its products and the characteristics of its client population, and therefore differs per institution. There is, however, a standard set of general ML scenarios and typologies that banks are expected to account for in their TM systems by DNB at least. Those flag unusual transactions that should always be investigated, without considering the context or customer specifically (interviewee 5).

Dutch financial institutions are mandated to conduct CDD reviews during the client onboarding process. Nevertheless, the intentions of a client are not static and may evolve over time. In addition, the client can be coerced into criminal activities or even be exploited as a money mule or ML front. Therefore, banks also have to continually monitor client behavior to ensure it aligns with established expectations and to detect any anomalies in transaction patterns (interviewee 5). To do so, banks can use an Expected Transaction Profile (ETP) of clients to serve as a benchmark to identify potential ML deviations from (interviewees 4, 5 and 7). Factors such as the client's risk classification (low, medium, high) are considered when establishing an ETP. Low-risk clients are allowed to be grouped in one peer group with one common ETP. As interviewees 6 and 7 note, effectively deploying an ETP on a per-client basis presents challenges. These challenges arise not only because an individual's financial situation can change, as in the case of 'someone could also just start earning more through a promotion or another job,' but also because the ETP, although validated at onboarding, may be constructed based on the client's narrative. Such self-reported narratives can be misleading or incomplete, leading to an ETP that seemingly validates even high-risk or dubious transactions, such as receiving funds from high-risk jurisdictions. For businesses, the chamber of commerce information such as activity codes, location and shareholders can be utilized (interviewees 2 and 6). Each risk level can have its own set of thresholds and unique indicators with varying absolute or relative values. While the use of ETP within Dutch AML efforts is increasing, interviewees highlighted that its effectively poses its own set of challenges and it has remained very limited within the retail domain up until now (NVB, 2023, interviewee 4).

Interviewees shed light on the workings of those rule-based Transaction Monitoring (TM) systems, highlighting that indicators in rule-based TM systems usually operate independently of each other (interviewees 1, 4 and 6). This is done to ensure the set lower bounds and to prevent indicators from cancelling out other indicators, potentially creating blind spots within the systems. To ensure the effective use of rule-based TM systems, tuning both static and dynamic thresholds used within rules is considered essential. As this tuning becomes more difficult as the number of parameters per rule increases, 'OR' statements are usually avoided in rule formulations (interviewees 2, 6 and 7). Alerts in rule-based TM systems are typically binary, activated when any indicator is triggered, and they usually do not provide further risk scores. There is a shared sentiment among interviewees that dynamic indicators are more effective.

When a transaction has been flagged, a bank officer will manually investigate whether the transaction is a false positive (not unusual) or true positive (unusual) and consequently should be reported as a hit to the compliance department. The officer might inquire of the client during the investigation for an explanation regarding the transaction. The compliance department will determine whether the transaction should be reported to the FIU and whether the particular client should be monitored more intensively, e.g., examining all transactions above a certain amount or conducting more frequent CDD reviews (interviewees 1 and 5). Conclusively, the section emphasizes the mandated procedures and the inherent challenges of effective TM within Dutch financial institutions.

4.3 Money Laundering Typologies

Transitioning to typologies, Table 9 presents the most significant ML typologies as identified by the interviewees. The typologies *Low to high*, *Geographical structuring/fan-out*, *Bipartite, Type change, Wallet, Profile inconsistency, Changing information* and *Circular/cycle* from Table 1 were

not mentioned by any interviewees and are therefore not included in Table 9. The typologies *Trade-based* and *Crypto* have been added based on the insight gathered from the interviews.

Table 9[20]. Most common money laundering typologies through bank transactions as mentioned by interviewees

Typology	Description	Referenced by interviewees
Structuring/ smurfing	Dividing large sums of money into smaller transactions to remain below mandatory reporting thresholds. This typology can be combined with almost all other typologies. With smurfing, usually structuring using multiple individuals is meant. This is a common tactic with individuals (money mules) where significant cashflows, often from activities like drug trafficking, are deposited in small amounts on personal accounts to avoid suspicion. Typically, amounts are kept under €10,000 to evade mandatory reporting requirements, but even transactions above this threshold must be consistent and justifiable. Funds frequently funnels through consistent accounts in a "pass-through" or "funneling" manner, ensuring that what comes in largely equals what goes out. Rapidly forwarded transactions and dormant accounts suddenly becoming active are suspicious indicators. Periodic CDD reviews are conducted on such accounts. Although there is an objective €10,000 limit for currency exchanges, no such limit exists for deposits, making this tactic particularly viable. Therefore, transaction amounts deliberately remaining just below reporting thresholds should also trigger alerts, as it indicates a deliberate attempt to evade detection. While banks should ideally determine periods for structured transactions, the criteria are often ambiguous, leaving room for various interpretations.	1, 2, 3, 5, 6, 7
Cash	Involves unusually high deposits or withdrawals, often executed either immediately after each other or within a short time frame. This typology is frequently combined with other ML methods, such as *smurfing*, particularly in typical scenarios involving drug dealers. In the retail sector, large cash transactions often hint at criminality or undeclared work. However, merely possessing cash is not inherently suspicious; it can become so when deposited or exchanged, as seen in drug-dealing scenarios involving smurfing to deposit funds. Monitoring typically involves comparing recent transactions to historical data. Large cash transactions are especially prevalent and are often linked to various forms of criminality. The risk associated with cash transactions varies by scale of criminal activity, being more significant for smaller-scale criminals as larger volumes often involve fronts for more extensive operations. Financial institutions may also consider regional variations and the scale of criminal activity when assessing risk. Periodic reviews are conducted to ascertain the legitimacy	1, 2, 3, 4, 5, 6, 7

[20]To prevent unintended adverse consequences related to money laundering from occurring, detailed numerical values are not published throughout this section.

	of large cash transactions, especially in countries with open economies where cash plays a significant role.	
Inactive/ dormant	Accounts that have been inactive for extended periods and then suddenly exhibit a surge in activity, frequently in the form of deposits, which may be combined with the *Cash* typology. Although dormant accounts are not intrinsically suspicious, maintaining an unused account for an extended duration is atypical. A sudden surge in activity, characterized by unusually high credit and debit transactions, serves as an indicator of potential ML. Such behavior is commonly observed in *smurf* accounts.	1, 2
International trade	Accounts that engage in transactions with countries different from those importing/exporting goods to or from, or providing services in. Notable shifts in international trade patterns, such as a regular supplier suddenly engaging with a new country, are red flags that warrant further investigation. Key questions include the rationale behind the shift in business focus to the new country and the security and credibility of that country.	1
Money mule /fan-in	Accounts registered under multiple individuals but controlled by a single party, frequently utilizing *passthrough* or *funneling* methods. An indicator of this typology is a balanced transaction flow where the volume of incoming funds closely aligns with outgoing funds. Rapid *passthrough* transactions are particularly suspicious. The activities may involve *cash* deposits and extend beyond criminal fund sourcing to include fraud or tax evasion. The strategy aims to obfuscate the traceability of funds, typically by moving funds from one account to multiple accounts before consolidating it back into a single account.	1, 2, 4, 5, 7
Pass through /funneling	Accounts exhibiting debit and credit transactions of closely matching amounts, indicative of a cover-up account. In the context of ML, this technique is often employed in conjunction with *money mules*. The strategy involves the rapid movement of funds, with swift forwarding transactions raising particular suspicion.	1, 2
Unrelated accounts	Transfers from or to unrelated accounts without reasonable justification. Anomalies such as a band receiving payment from an unexpected location following an overseas concert, or a student whose expenses are entirely covered by an unrelated third party, warrant scrutiny. This typology may intersect with other methods like *money mules* when *structuring* is executed within a network.	1, 7
Currency exchange	Substantial and continuous purchase transactions for foreign currency (in cash) or cryptocurrency are deemed unusual, as it can often be a method to obscure the origins of illicit funds.	1
High-Risk Jurisdictions	Transfers involving high-risk jurisdictions, including tax havens and countries frequently linked to tax evasion and schemes like Mexico's black market peso exchange, are subject to heightened scrutiny. The risk increases with payments facilitated through transfer services. Regulatory entities like the World Bank, UN, and EC provide lists of high-risk countries that guide both outbound and inbound	1, 2, 3, 5, 6, 7

	TM. CDD is typically mandated for transactions involving these regions. The typology is especially pertinent for retail customers, who typically engage in transactions within the Netherlands and Europe. Transactions involving high-risk countries outside this geographical scope are considered unusual unless adequately justified. The risk profile varies depending on the frequency and nature of transactions, making it crucial for financial institutions to adapt their monitoring strategies accordingly.	
Laundromat	Involves the use of front companies that default on fabricated loans, which are then authenticated by corrupt judicial authorities, allowing for payments from third parties. This method is often interlinked with *High-Risk Jurisdictions* and schemes like *Trade-based*, particularly in relation to drug money. Notably, this typology has been observed in cases involving shells, where funds were *funneled* through intricate networks. This approach often employs *structuring* to conceal relationships, especially in trade-based schemes.	1, 6, 7
Trade-based	Involves the manipulation of trade invoices—either under-invoicing, over-invoicing, or double invoicing—to facilitate illicit financial flows between organizations or jurisdictions. From a banking standpoint, this typology is particularly challenging to detect and can manifest in two primary forms: Documentary Trade and Open Account Trade. In the case of Documentary Trade, banks are involved in the transactions providing financial instruments like guarantees, thereby gaining access to transaction details such as invoices, allowing for a more informed assessment of the transaction's legitimacy. Conversely, Open Account Trade poses a greater challenge as banks facilitate the payment but have limited information about the transaction itself. This typology is frequently used to convert money from one currency to another, especially from high-risk jurisdictions, through a series of seemingly plausible transactions that may involve multiple "intermediary" banks or countries. Each link in this transactional chain has the potential to act as a conduit for ML. Given the typology's complexity and the possibility of involving both legitimate and illegitimate parties, vigilant monitoring of both sides of a transaction becomes imperative for banks. Network analysis is especially efficacious when a financial institution serves both parties involved in a transaction.	4, 6
Crypto	Involves the conversion of cryptocurrency to fiat currency, often through complex transactions designed to obscure the origin of funds. The process is a crucial point of concern as it presents a juncture where the traditional financial system intersects with decentralized cryptographic assets, meriting stringent regulatory and monitoring scrutiny. Given the obligatory involvement of banks during the conversion process, this typology adds a layer of complexity and urgency to contemporary AML efforts and transaction monitoring protocols.	5

4.3.1 Selected Money Laundering Typologies

As the information presented in Table 9 indicates, the most common transactional ML typologies in the Netherlands, as highlighted by the interviewees, are *Cash*, S*tructuring/smurfing* and *High-Risk Jurisdictions*. The focus on these three typologies is a result of their frequent mention and perceived prominence in the interviews, establishing their importance within the Dutch context.

4.4 Indicators of Money Laundering Typologies

With the typologies established, the next logical step is to discuss the indicators that can flag these typologies. Table 10 presents these indicators that can be used to generate alerts for the subset of typologies as described directly above. The indicators *Sector salaries, Sector high-risk* and *Account duration* from Table 3 were not mentioned by any interviewees and are therefore not included in Table 10. The indicators *Number roundness*, *Transaction description*, *Structure of banknotes* and *Growth of company* have been added based on the insight gathered from the interviews.

Using predefined thresholds, financial institutions try to delineate the expected transactional behavior of their customers. While such thresholds are relatively straightforward for most retail customers, they become increasingly complex for small-scale international retailers. This inherent variability complicates the task of devising a universally applicable solution for TM across diverse customer segments (Interviewee 6). Within the framework of these indicators, both static and dynamic thresholds are used within rule-based TM systems at Dutch banks (interviewees 1 and 2). This can range from "cash deposits above amount X" to "more than X% of your average account amount". The amount and type of thresholds depend on the typology and can differ per the anticipated risk category of a client or product (Interviewees 1, 2 and 3). In addition, rules can be set with conditional parameters, e.g., no alert is generated when the transaction amount is too small. The number of rules within an "average" TM system varies greatly by bank and by department, e.g., retail, wholesale or private banking (interviewees 1 and 6). While multiple rules can be set, more rules are not always more effective, as even a single alert for an unusual transaction already warrants further investigation (interviewee 6).

Table 10[21]. Indicators of common money laundering typologies as mentioned by interviewees

Indicator	Description	Referenced by interviewees
Type	Type of transaction. Can be a 'normal' bank transfer, cash deposit or withdrawal, transfer via another payment provider, loan payment, or cheque.	1, 2, 3, 4, 6, 7
Amount	The amount of the transaction in relation to the static (e.g., student account) or dynamic threshold.	1, 2, 3, 4, 5, 7
Frequency	The frequency of the transaction (potentially of a similar amount) in relation to the static or dynamic frequency threshold.	4, 7
Volatility	Rapid and relatively large changes in the account balance over a brief period, characterized by significant percentages of inflow or outflow.	1, 2, 3, 4, 7
International exposure	Percentage of in or outflow to or from abroad in relation to static or dynamic thresholds.	3, 6

[21] To prevent unintended adverse consequences related to money laundering from occurring, detailed numerical values are not published throughout this section.

International high-risk	Transactions involving or originating from high-risk countries.	3, 4, 5, 6, 7
Account type	Personal (e.g., student or normal) or business account.	1, 2
Number roundness	Degree to which the transaction amount is a "round" number.	1, 4
Transaction description	Relevance and patterns in the transaction's text description.	2
Structure of banknotes	Denomination and structure of cash in deposits.	4
Growth of company	Unnatural growth rates, incongruent with economic climate and regional trends.	4

4.4.1 Selected Indicators of Money Laundering Typologies

It should be noted that although all indicators outlined in Table 3 could ideally be employed to generate alerts for the subset of typologies practical limitations exist. As outlined in Table 6, the IT-AML dataset does not contain any information, such as the location, sector or risk classification, about the account holders. This prevents the creation of ETPs, which renders the use of the *Growth of company* and *Account type* indicator impossible. In addition, as no transaction description or additional information is provided on *Cash* transactions, the *Transaction description* or *Structure of banknotes* indicators can also not be leveraged. However, the transaction data does contain information about the date and time, involved accounts, amounts, currencies and the type of transaction. This enables use of all the remaining indicators presented in Table 10.

Drawing upon the insights gained from the interviews, three or two indicators were selected for each typology under study. Table 11 presents, per each respective typology, the indicators employed to formulate rules aimed at detecting instances of that particular typology. The logic of flagging transactions is that the indicators work independently from each other. In other words, a particular transaction can be flagged multiple times by different indicators independently within the same typology.

Table 11. Indicators of selected money laundering typologies

Typology	Indicators	Indicator Type	Indicator Logic
Cash	Type	Categorical	*Type* == Cash
	Amount	Static	*Amount* ≤ value X
Structuring/ smurfing	Amount	Static	*Amount* ≤ value X
		Dynamic	*Amount* ≥ more than X times average transaction amount for account
	Frequency	Dynamic	*Transaction count per day* ≥ X% above the average daily transaction count for account
	Volatility	Dynamic	*Daily net flow* ≥ more than X standard deviations above average net flow for account
High-Risk Jurisdictions	International high-risk	Categorical	*Currency* == high-risk currency proxy

	International exposure	Dynamic	*Transactions non-dominant currency* ≥ more than X% of transactions in account's non-dominant currency
	Amount	Static	*Amount* ≤ value X

4.5 Chapter Conclusions

- The most common transactional ML typologies in the Netherlands are *Cash*, S*tructuring/smurfing* and *High-Risk Jurisdictions*.
- There is a shared sentiment that dynamic indicators are more effective than static ones.
- For the *Cash* typology, the *Type* and *Amount* indicators will be used.
- For the *Structuring/smurfing* typology, the *Amount*, *Frequency* and *Volatility* indicators will be used.
- For the *High-Risk Jurisdictions* typology, the *International high-risk*, I*nternational exposure* and *Amount* indicators will be used.

5. Data and Experiments

This chapter presents how the typologies and indicators as selected were handled in the data and provides information on the experimental setups used for training and evaluating the classifiers.

5.1 Data

5.1.1 Data Preprocessing

Since large amounts are usually not indicative of ML, as discussed previously in Chapter 4, transactions with a top 10% amount value have been removed, leading to the distribution as presented in Figure 8. This caused the removal of 17.225.065 transactions from the high ML rate and 16.880.162 transactions from the low ML rate dataset.

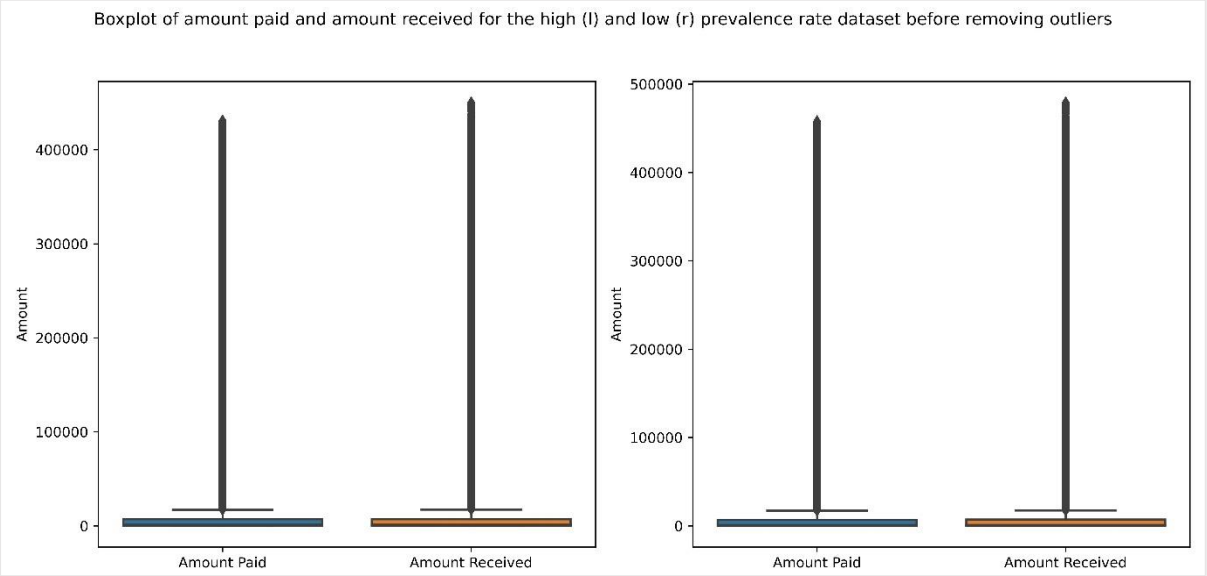

Figure 8. Boxplots of amount variables in both datasets after removing outliers.

For training the models and making predictions, the Amount Received column has been removed to reduce multicollinearity because it was, obviously, highly correlated to the amount paid column. To obtain betters fitted models, categorical variables, such as bank names or currencies, have been transformed into numerical representations. One-hot encoding has been applied to the columns 'Receiving Currency', 'Payment Currency' and 'Payment Format'. Robust scaling was applied to the numerical 'Amount Paid' column due to the positive skewness and large outliers (Scikit-learn developers, n.d.-d). Due to the high class imbalance in the (train) datasets, we used Repeated Stratified K-Fold for cross-validation to ensure that each fold adequately represents the minimum level of ML transactions present in the data (Brownlee, 2020a).

5.1.2 Training and Testing Data

The train/validate/test split has been based on time, using the "Timestamp" of each transaction in the dataset, as per Jullum et al. (2020). This time-based split is crucial in the context of TM, as it more accurately mirrors the temporal dynamics in transaction patterns. In addition, this facilitated an out-of-time validation of model performance, as if the model would have been enabled at a certain point in time (Zhang & Trubey, 2019). We used a 60%/20%/20% train, validate and test split as this was advised by the developers of the IT-AML datasets to be most beneficial (Altman, 2023). This meant that, from the 97 days of the total timespan of the datasets, the first 57 days were used for training, the next 20 days for validating and the last 20 days for testing. This is the same as Jullum et al. (2020) did, as they used two months of transaction history. The number of transactions in the various splits are presented in Table 12.

Table 12. Number of transactions and prevalence rate after the train/validate/test split

	High prevalence rate		Low prevalence rate	
Dataset	# of transactions	Prevalence rate	# of transactions	Prevalence rate
Train	90,646,164 (58%)	0,10%	88.843.454 (58%)	0,05%
Validate	31,850,904 (21%)	0,13%	31.202.441 (21%)	0,06%
Test	32,528,821 (21%)	0,13%	31.875.978 (21%)	0,06%
Total	**155,025,889**		**151.921.873**	

5.2 Typologies and Experiments

As stated in 4.5 Chapter Conclusions, the *Cash*, S*tructuring/smurfing* and *High-Risk Jurisdictions* have been used throughout this study. This section discusses how the indicators for the various typologies and indicators were handled. Once again, to prevent unintended adverse consequences related to ML from occurring, detailed numerical values are not published throughout this section. For all three typologies, all transactions below a certain minimum value were excluded as interviewees indicated that ML using small amounts is not considered feasible. For this end, all amounts were converted using the average exchange rate over the total timespan. All values reported for the typologies below are for the 'all banks' perspective.

5.2.1 Cash Typology

For the *Cash* typology, the *Amount* and *Type* indicators were employed. The Cash payment format type was the primary indicator used for flagging transactions. This indicator flagged 6,733,824 transactions with a ML rate of 0.02% in the high ML rate dataset and 6,627,245 transactions with a ML rate of 0.02% in the low ML rate dataset. Next to that, the minimum transaction amount rule removed 19.505.841 transactions from the high ML rate and 19.155.000 transactions from the low ML rate dataset. The overall ML rate for the flagged transactions was 0.02% in the high ML rate cash train dataset and 0.02% in the low ML rate cash train dataset. We note that these rates are very limited in comparison to the rates in the train sets of the other two typologies.

5.2.2 Structuring/smurfing Typology

For this typology, the *Amount*, *Frequency* and *Volatility* indicators have been used. The *Amount* rule (both minimum transaction amount value and those exceeding thresholds) flagged 2,931,594 transactions with a ML rate of 0.42% in the high ML rate dataset and 2,868,359 transactions with a ML rate of 0.14% in the low ML rate dataset. The Volatility indicator was used by flagging instances where sudden and irregular fluctuations in transaction values deviate significantly from the norm, potentially indicating attempts to obscure illicit funds through rapid and unpredictable movements. As the account balance was not known, the volatility has been determined using cashflows occurring within the timespan of the dataset. This indicator flagged 2,619,422 transactions with a ML rate of 0.53% in the high ML rate dataset and 2,588,933 transactions with a ML rate of 0.08% in the low ML rate dataset. The Frequency indicator has been employed to detect ML transactions of the structuring/smurfing typology by identifying a high volume of small, repetitive transactions that may be indicative of an attempt to break down a large sum of illicit funds into smaller, less conspicuous amounts to evade detection. This indicator flagged 1,072,256 transactions with a ML rate of 0.40% in the high ML rate dataset and 1,042,511 transactions with a ML rate of 0.11% in the low ML rate dataset. The overall ML rate for the flagged transactions was 0.46% in the high ML rate high-risk jurisdictions train dataset and 0.12% in the low ML rate high-risk jurisdictions train dataset.

5.2.3 High-Risk Jurisdictions Typology

For this typology, the *Amount*, *International high-risk* and *International exposure* indicators were used. The Amount indicator removed 19,505,841 transactions from the high ML rate and 19,155,000 from the low ML rate dataset because they remained below the minimum transaction value. The currencies in the datasets have been used as proxies for which countries the money

would go to. The countries China, Russia, Mexico and Brazil are considered high-risk (International Centre for Asset Recovery (ICAR), 2022; KnowYourCountry, n.d.-c, n.d.-a, n.d.-b; Statista, 2023a, 2023b). The International high-risk indicator flagged 10,826,692 transactions with a ML rate of 0,08% in the high ML rate and 9,508,219 transactions with a ML rate of 0,06% in the low ML rate dataset. In addition, all transactions of the Bitcoin currency and payment format are assumed to be high-risk as well. Unusual transactions were also identified using the *International exposure* indicator. More specifically, a transaction was flagged if it exceeded a predetermined percentage of the average for transaction amounts denominated in currencies divergent from an account's 'standard' currency. This rule flagged 4,056,461 transactions with a ML rate of 0.60% in the high ML rate and 12,982,762 transactions with a ML rate of 0.10% in the low ML rate dataset. The overall ML rate for the flagged transactions was 0.22% in the high ML rate high-risk jurisdictions train dataset and 0.07% in the low ML rate high-risk jurisdictions train dataset.

5.2.4 Experiments

This section presents the various experiments conducted on the dataset. To enhance the study's generalizability, the FNR has been empirically calculated using ground-truth data and estimated using predictive models across two distinct datasets, as presented earlier in 3.3 Synthetic Data. As Dutch banks are increasingly collaborating to combine transaction data in one system, called 'Transactie Monitoring Nederland' (Transactie Monitoring Nederland, n.d.). Therefore, we will conduct the experiments from both the perspective (information position) of one bank as well as from all banks. For the perspective of one bank, we use bank 70, as this is the most predominant bank in the datasets, as seen in Figures 8 and 9 in Explorative Data Analysis and data preparation. To this end, we will filter all transactions not originating from or going towards bank 70. Since we are training seventeen distinct models per scenario, the total number of combinations is 204 (2*2*3*17). Due to computational constraints, we have randomly sampled 20% of the training and test dataset before fitting and predicting with every model.

Experiment A. High ML Prevalence Rate from the All Banks Perspective

- *Cash* typology on high ML prevalence rate dataset from all banks perspective.
- S*tructuring/smurfing* typology on high ML prevalence rate dataset from all banks perspective
- *High-Risk Jurisdictions* typology on high ML prevalence rate dataset from all banks perspective

Experiment B. High ML Prevalence Rate from the One Bank Perspective

- *Cash* typology on high ML prevalence rate dataset from one bank perspective
- S*tructuring/smurfing* typology on high ML prevalence rate dataset from one bank perspective
- *High-Risk Jurisdictions* typology on high ML prevalence rate dataset from one bank perspective

Experiment C. Low Prevalence Rate from the All Banks Perspective

- *Cash* typology on low ML prevalence rate dataset from all banks perspective
- S*tructuring/smurfing* typology on low ML prevalence rate dataset from all banks perspective
- *High-Risk Jurisdictions* typology on low ML prevalence rate dataset from all banks perspective

Experiment D. Low Prevalence Rate from the One Bank Perspective

- *Cash* typology on low ML prevalence rate dataset from one bank perspective
- S*tructuring/smurfing* typology on low ML prevalence rate dataset from one bank perspective
- *High-Risk Jurisdictions* typology on low ML prevalence rate dataset from one bank perspective

5.2.5 Experimental Setup

As aforementioned in 3.2.2 Machine Learning Classifiers, the default implementations of the classifiers logistic regression, decision tree and random forest are considered as baseline models. Next to those, we employ a range of these and additional models with cost-sensitive learning and downsampling approaches to address the high class imbalance in our datasets.

Cost-sensitive Learning

Cost-sensitive learning helps to increase model performance by assigning a higher penalty to misclassifying the minority class, incentivizing the models to focus on accurately detecting ML instances. This approach is particularly beneficial for dealing with the low ML prevalence dataset, as it presents more significant predictive challenges due to less frequent ML activities, making the discovery of patterns more complex (Altman et al., 2023). To implement cost-sensitive learning, we leveraged the class weight parameter for the following models: Logistic Regression, Decision Tree, Random Forest, Balanced Random Forest and AdaBoost with Decision Trees. We used the 'balanced' option for these models, which adjusts weights inversely proportional to class frequencies in the input data, therefore prioritizing minority class predictions in the loss function. As the Random Forest and Balanced Random Forest models use bootstrapping for the training of different trees, they also offered a 'balanced_subsample' option for this parameter, which we also implemented for a version of both models. This also adjusts weights inversely proportional to class frequencies, but then on a per-bootstrap sample basis.

Downsampling

Next to cost-sensitive learning, we also leveraged a downsampling approach to increase model performance, which effectively manages the significant volume disproportion between the 'ML and 'non-ML' class. These models, Balanced Random Forest and RUSBoost, are designed to downsample the majority class in the training dataset, thereby creating a more balanced class distribution. By doing so, they ensure a more equal representation of classes during the learning process, which is particularly crucial in our context where the ML cases are substantially less prevalent than non-ML transactions. The downsampling approach helps in mitigating the model's bias towards the majority class, enhancing its sensitivity to the minority class, and thus improving the detection of potential ML transactions (Hasanin & Khoshgoftaar, 2018). The specific downsampling strategy where implemented is 'auto' without replacement, a choice made to randomize the downsampling process and ensure equal class sizes. This entails dropping the majority 'non-ML' class until an equal split has been met, therefore completely diminishing the class imbalance. This 50/50 class representation makes the model less able to overfit on the majority class, thereby improving the detection capabilities for the minority class. To provide a balanced perspective, it should be noted that while this approach is effective in reducing bias, it may involve the loss of some informative instances from the majority class, potentially impacting the model's overall learning capacity.

Metrics for Evaluation

We need to define the metrics used to compare the predictions from the machine learning classifiers with the performance of the rule-based TM systems. The subject of study is transactions misclassified as 'not unusual' by rule-based TM systems—known as *false negatives*. It is worth noting that rule-based TM systems in banking institutions typically flag only a minimal subset of transactions for manual review. Consequently, the true class label for the vast majority of transactions deemed usual remains unverified, yielding a confusion matrix as depicted in Table 13.

Table 13. Confusion matrix for real-world AML TM systems.

		Actual class	
		True (unusual)	**False (usual)**
	True (unusual)	True Positive	False Positive

Predicted class	**False (usual)**	False Negative + True Negative

However, since we have access to ground-truth synthetic datasets wherein transactions are unambiguously labelled as instances of ML or not, we also have knowledge of the true class labels, as outlined in Table 14. This enables the calculation of the False Negative Rate (FNR), also known as the miss rate (Herzog et al., 2017). The FNR serves as an essential metric in our study, given our interest in identifying false negatives. It quantifies the rate at which actual unusual transactions are erroneously not flagged by the system. The FNR can be calculated by dividing the number of false negatives by the total number of actual positives: $FNR = FN/P$ ("Confusion Matrix," 2023).

Table 14. Confusion matrix of TM systems using synthetic datasets.

		Actual class	
		True (unusual)	**False (usual)**
Predicted class	**True (unusual)**	True Positive	False Positive
	False (usual)	False Negative	True Negative

Finally, we use the Area Under the Precision-Recall Curve (AUPRC) as the primary metric to determine model performance and the best hyperparameter setup as this metric is specifically tailored towards capturing a classifiers ability to correctly predict the minority class. To calculate AUPRC, we plot precision (true positives divided by all positive predictions) against recall (true positives divided by actual positives) across various thresholds, forming a curve. The AUPRC, ranging from 0 to 1, represents the area under this curve, quantifying the model's overall ability to distinguish the minority class. A score of 1, indicating perfect precision and recall, reflects enhanced model performance in detecting these instances. We note that the baseline of the AUPRC is equal to the fraction of positives, as calculated by dividing the number of positive examples by the total number of examples (Saito & Rehmsmeier, 2015). For the high ML prevalence dataset this is 0.13% (1/807) and for the low ML prevalence dataset this is 0.06% (1/1,750).

Delving into the robustness of the predictions, the Matthews Correlation Coefficient (MCC) is reported as the secondary metric to determine model performance in general, which is in line with Chicco & Jurman (2020) who advocate for the robustness of MCC in binary classification model evaluation. Offering a more balanced and reliable approach than straightforward accuracy in scenarios with imbalanced class distributions like ours, the MCC stands out by considering both true and false positives as well as negatives in its computation. In contexts, where false negatives (unidentified ML transactions) are of particular interest, the MCC's sensitivity to both types of classification errors becomes crucial. The MCC ranges from -1 to +1, where +1 represents a perfect prediction, 0 is no better than random prediction, and -1 indicates total disagreement between prediction and true class.

Model Parameter Optimization

The best performing model for every experiment/typology combination was selected for further optimization through hyperparameter tuning, a process where we adjust the model's parameters to improve its performance. To manage the computationally expensive nature of hyperparameter tuning, we employed random search cross-validation. This approach, while less exhaustive than a full grid search, has been shown to yield comparably effective results (Alice, 2016). Repeated Stratified K-Fold cross-validation was applied using the training set (Brownlee, 2020a). Models were tuned with a range of hyperparameters tailored to their specific characteristics, as presented in Appendix C. Hyperparameter tuning. The number of random samples for the tuning

ranged from 10 to 50, dictated by the model's training time on each dataset. For final model evaluations, we used the set of hyperparameters that achieved the highest AUPRC score.

Threshold Tuning

Having retrained the best model for every experiment/typology combination using the optimal hyperparameter settings, we attempted to further improve the classification performance by applying threshold tuning (Brownlee, 2020b). This process involves adjusting the decision threshold, which is the point at which a model decides between different classification outcomes. Typically, a default threshold of 0.5 is used for binary classification, but this may not be optimal for all situations, especially in imbalanced datasets. To find the optimal threshold, we explored a range between 0 and 1, using increments of 0.05. The optimal threshold was identified by maximizing the MCC across these different thresholds. By fine-tuning this threshold, we aimed to optimize the balance between false positives and false negatives, thereby enhancing the overall effectiveness and precision of our models in making predictions. It is important to note that adjusting the threshold does not change the AUPRC because this metric evaluates the model's performance across all possible thresholds, making it invariant to any single threshold change. However, it can significantly improve the MCC, which is why we used this metric for tuning.

6. Results

This chapter presents the results of the experiments. As indicated above in Metrics for Evaluation, the performance of the classification models is evaluated based on the Area Under the Precision-Recall Curve (AUPRC) and Matthews Correlation Coefficient (MCC) metrics. As described above in Model Parameter Optimization, we have selected only the best performing model for every experiment/typology combination for hyperparameter and threshold tuning and making final predictions.

6.1 Experiment A. High ML Prevalence Rate from the All Banks Perspective

For the first experiment, the models were trained on a typology train set with the high ML rate and with the data of all banks. Figure 9 presents a comparative analysis of various machine learning models' performance, measured by the AUPRC. As we can see, the best-performing model varies across different typologies, indicating a need for a nuanced approach when selecting models for different ML scenarios. Ensemble methods, particularly those implementing balancing techniques, show substantial efficacy, suggesting that balancing strategies are beneficial in managing class imbalances typical in ML detection. Although the top ranking order is very similar, AdaBoost is a notably better performing model for the Cash typology than for high-risk jurisdictions and structuring typologies. Logistic Regression, both as default and balanced, appears to perform the worst for both high-risk jurisdictions and structuring, but not cash, indicating its limitations in handling complex patterns associated with ML activities within the former typologies. While the AUPRC scores are in a similar (already low) range for the high-risk jurisdictions and structuring typologies, the models seem to fail dramatically for the cash typology as the AUPRC reaches the ML prevalence rate, indicating performance on par with random guessing.

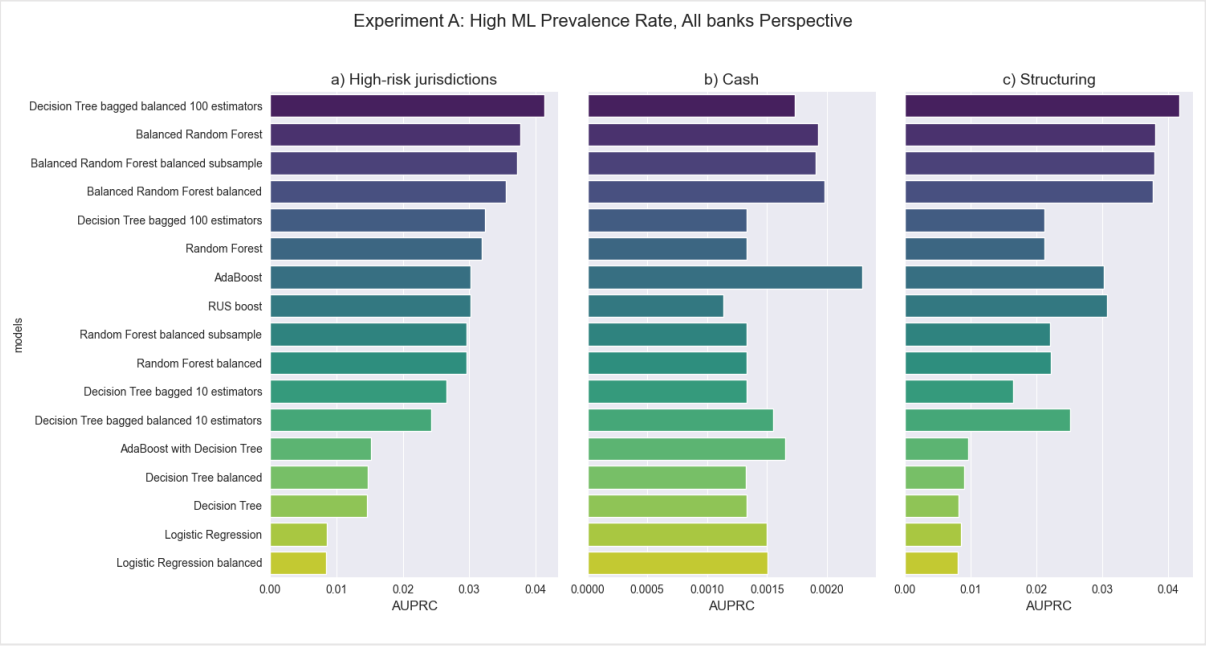

Figure 9. AUPRC scores for every model per typology in experiment A

We now compare the predictions of the best performing classifier for every typology with the actual ML cases present in the data. Figure 10 shows the results of the classifiers ability to detect and predict ML activities, depicted through the distribution of true positives, false positives and false negatives for each model/typology combination. For the structuring and especially the cash typology, the number of false positives is extremely large, indicating a very conservative prediction model. However, the number of true positives (the overlap), is very minimal when compared to the total number of positives, indicating poor model performance. This indicates

that while the model does predict ML instances, it does so with substantial inaccuracy, leading to many incorrect predictions. For the high-risk jurisdictions typology, there is a significant number of false positives as well. Despite this, there is a notable overlap between true and predicted cases, suggesting a measurable degree of accuracy in the model's predictions. In summary, across all typologies, the models display varying degrees of predictive capability, with none achieving high precision.

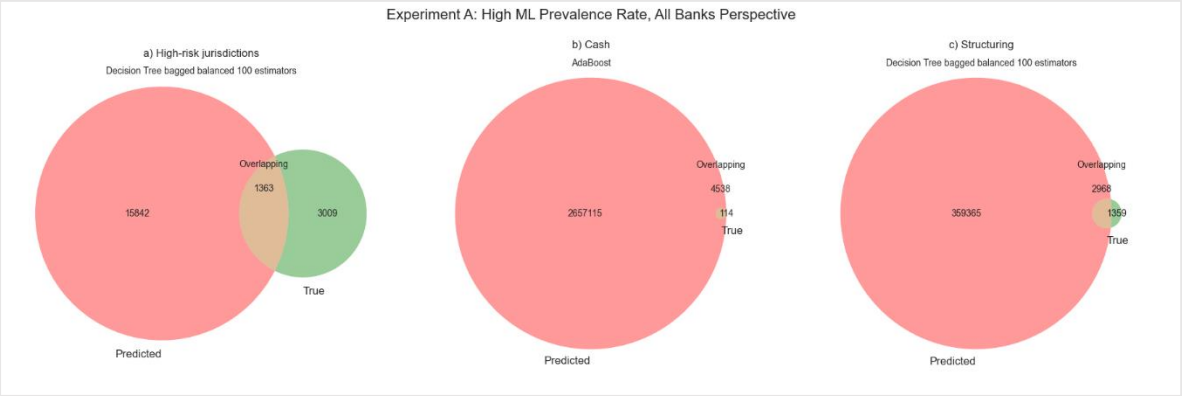

Figure 10. True and predicted number of ML transactions for the high ML rate data and the perspective of all banks

As we can see in Table 15, the difference between the actual and the Predicted FNR is very substantial for the cash and structuring typologies. However, the difference for the high-risk jurisdictions typology is very limited. However, the low AUPRC and MCC scores for this typology seem to suggest that this just is a coincidence and not an accurate correct estimation. The low AUPRC and MCC for the cash typology indicate that the model is not performing much better than random guessing. For the high-risk jurisdictions and structuring typologies, the model has learned to identify ML transactions somewhat better, but both metrics can still be considered low.

Table 15. Results of the balanced bagging with Decision Trees with 100 estimators classifier for all banks on the high prevalence dataset

	High-risk jurisdictions	**Cash**	**Structuring**
Actual FNR	0.731	0.990	0.924
Predicted FNR	0.688	0.025	0.313
AUPRC	0.053	0.002	0.019
MCC	0.155	0.015	0.067

6.2 Experiment B. High ML Prevalence Rate from the One Bank Perspective

Figure 11 shows a shift in classifier performance when the perspective changes to a single bank. The disparity in AUPRC performance between most models highly differs per typology, indicating that certain models excel for all typologies whereas others fail greatly depending on which typology its detection is aimed. The Balanced Random Forest balanced model leads, particularly for the high-risk jurisdictions typology, closely followed by AdaBoost. This could indicate that from a single bank's perspective, the class imbalance problem is less severe, or the data patterns are more uniform, allowing simpler models to perform relatively well. While the Balanced Random Forest implementations with the various balancing measures again perform well, the top performer from experiment A, balanced bagging with 100 decision trees, now drops in ranking significantly, together with the other balanced bagging models. The normal Random Forest models without downsampling perform poorer than their equivalent with downsampling, indicating that this balancing measure helps to improve model performance greatly.

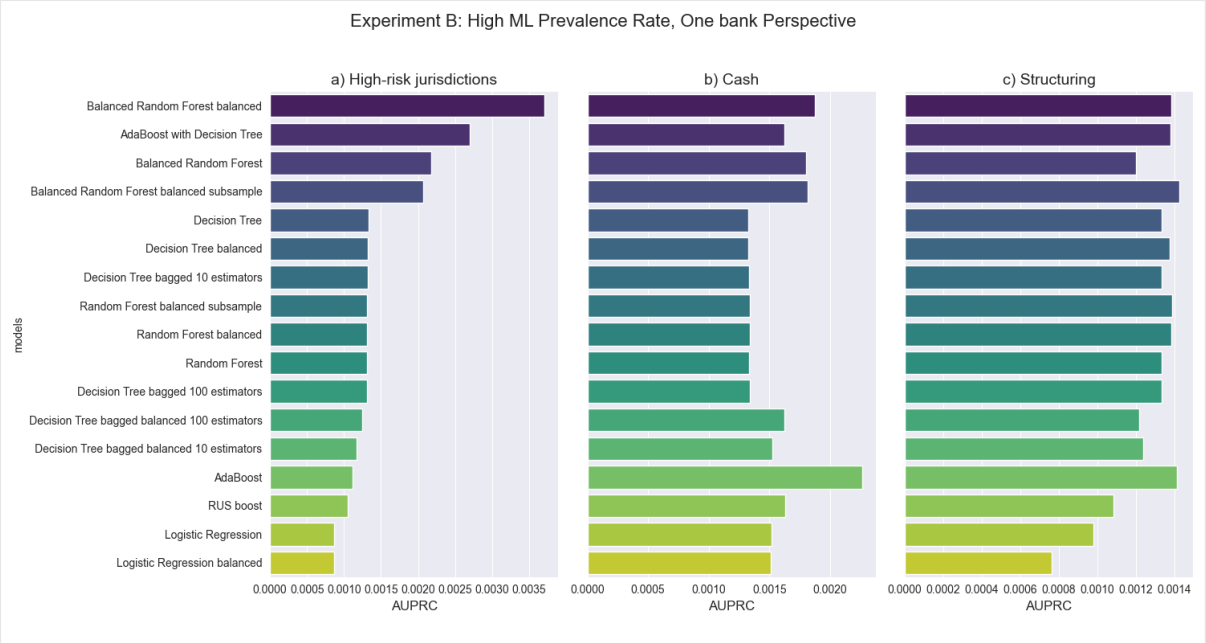

Figure 11. AUPRC scores for every model per typology in experiment B

The results as presented in Figure 12 paint a different, but worse, picture than those in experiment A. The classifiers once again perform very poorly. We can see that the number of false positives is very great for the high-risk jurisdictions and structuring typologies, indicating that the Balanced Random Forest model in those cases is once again conservative. Thus, the classifier has low performance although many actual ML transactions are correctly predicted, indicating low precision. AdaBoost for the cash typology predicts ML a relatively limited number of times but has almost no correct overlapping true positives. The model might not be sufficiently sensitive to actual ML activities, potentially leading to a high FNR in real-world applications.

Overall, the Venn diagrams in Figure 12 underscore the challenge of achieving a balance between sensitivity and specificity in ML detection models. While a conservative model may minimize the risk of missing actual ML transactions, the high number of false positives could lead to inefficient allocation of investigative resources and operational inefficiencies in a real banking context.

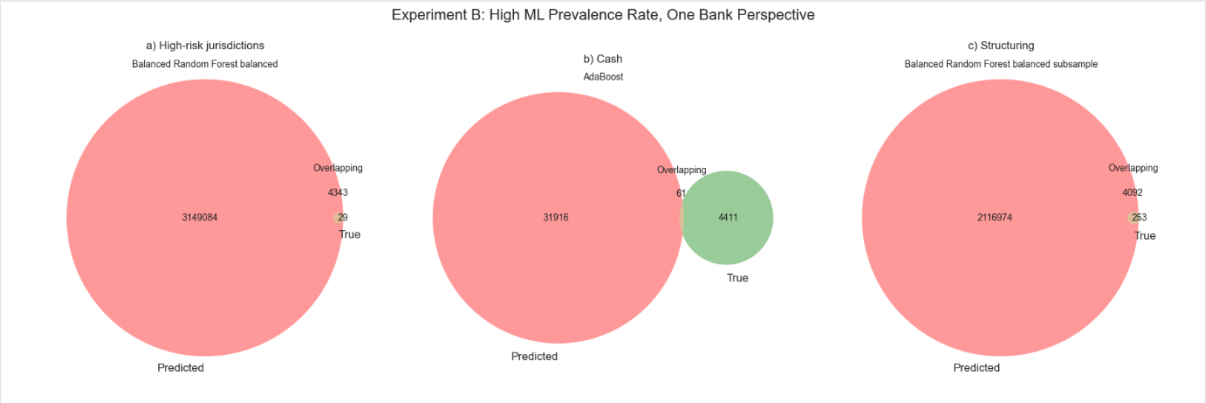

Figure 12. True and predicted number of ML transactions for the high ML rate data and the perspective of one bank

Table 16 highlights a pronounced gap between Actual and Predicted FNR for the typologies high-risk jurisdictions and structuring, indicating many actual ML transactions are missed. For the cash typology, the actual and predicted FNR are not that far off. However, just as for Experiment A, this seems rather a coincidence. Especially given that the low AUPRC values across the board signal poor model precision and recall, with the model's performance barely surpassing random chance, as also suggested by the near-zero MCC values. This suggests the limited effectiveness of all

classifiers in accurately detecting ML activities in the low rate dataset from the perspective of one bank.

Table 16. Results of the Balanced Random Forest with balanced subsample classifier for one bank on the high prevalence dataset

	High-risk jurisdictions	Cash	Structuring
Actual FNR	0.892	0.908	0.959
Predicted FNR	0.007	0.986	0.058
AUPRC	0.001	0.001	0.001
MCC	0.005	0.001	0.022

6.3 Experiment C. Low ML Prevalence Rate from the All Banks Perspective

In Figure 13, we observe classifier performance in a low ML rate prevalence scenario from an all banks perspective. As anticipated, there is a general decline in performance across all classifiers compared to the high ML rate prevalence of Experiment A, which is expected due to the relatively reduced frequency of ML cases, which inherently complicates the detection process. With the same models in the top 4 as for experiment A, the balanced models seem to maintain a relative advantage, with the bagged balancing approach using Decision Trees and the Balanced Random Forest classifiers appearing prominently at the top. This suggests that methods designed to account for class imbalance are advantageous, even when the overall prevalence rate is low. In addition, it shows that the same ranking of performance maintains when varying the ML rate. Noteworthy is the poor performance of Decision Tree and Random Forest models, which are less capable in this context, likely due to their inability to adequately differentiate between the majority class of non-ML transactions and the sparse ML cases. The uniformly lower AUPRC values across all typologies and models highlight the intrinsic challenges of detecting ML activities in a low prevalence context and underscore the necessity for employing sophisticated models that are resilient to such class imbalances.

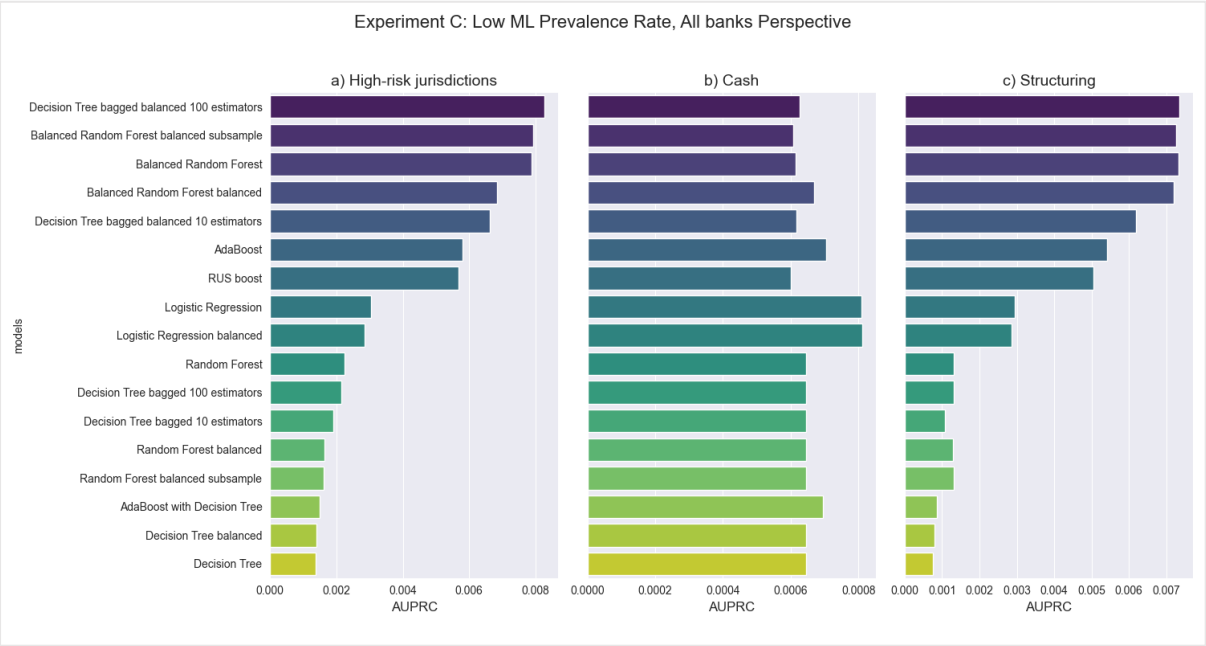

Figure 13. AUPRC scores for every model per typology in experiment C

Figure 14 displays the results from a bagged decision tree classifier with 100 estimators for the high-risk jurisdictions and structuring typologies. The cash typology model is Logistic Regression with cost-sensitive learning. Similar to experiment B, the classifiers predict a high volume of false positives compared to true positives across all typologies, indicating an inclination towards over-predicting non-ML transactions as ML. The small overlap between predicted and actual ML

transactions signifies the classifiers low precision in correctly identifying ML activities, suggesting a high false positive rate which could be problematic in operational banking applications. Together, these results confirm the difficulty of accurately detecting ML activities in a low prevalence environment and highlight the importance of developing classifiers that are not only sensitive to true ML transactions but also specific enough to avoid excessive false alarms.

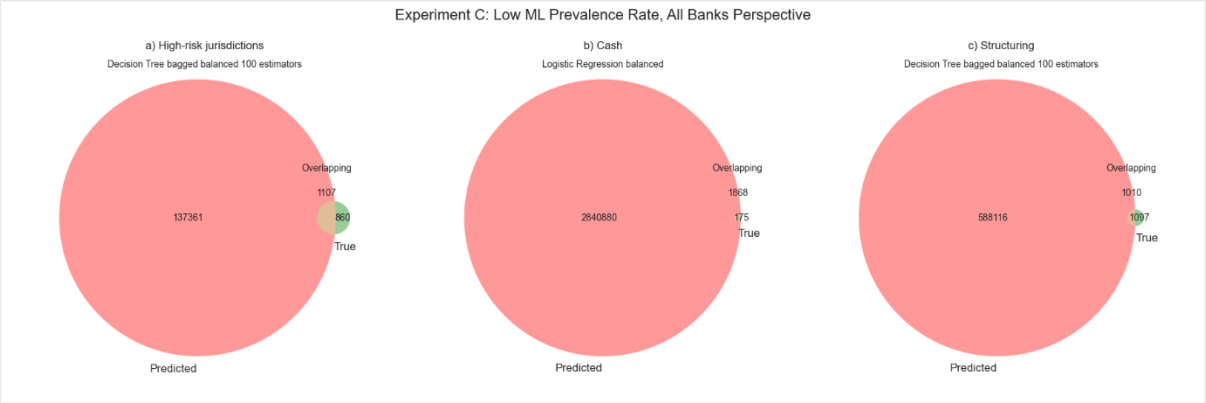

Figure 14. True and predicted number of ML transactions for the low ML rate data and the perspective of all banks

The data in Table 17 show a significant disparity between the Actual and Predicted FNR for all ML typologies. The low AUPRC and MCC values across all typologies suggest the model's limited effectiveness, performing only slightly better than random chance, especially for cash and structuring transactions. When comparing the scores to the high ML rate as presented in Table 15, we can clearly see that the AUPRC and MCC scores for all typologies are significantly lower, indicating that the classifier has more trouble differentiating between non-ML and ML transactions.

Table 17. Results of the balanced bagging with Decision Trees with 100 estimators classifier for all banks on the low prevalence dataset

	High-risk jurisdictions	**Cash**	**Structuring**
Actual FNR	0.825	0.979	0.961
Predicted FNR	0.437	0.086	0.521
AUPRC	0.008	0.001	0.002
MCC	0.063	0.002	0.020

6.4 Experiment D. Low ML Prevalence Rate from the One Bank Perspective

Figure 15 presents the AUPRC considering a single bank's data with a low prevalence of ML activities. The AUPRC values across the board are modest, reflecting the challenge of detecting rare events within a single institution's transactional dataset. This single-bank scenario brings a shift in the relative performance of the classifiers; the Logistic Regression, both as default and balanced, model exhibits a significant improvement in ranking compared to the other experiments, now performing comparably to its counterparts and even rivaling more complex models. Notably, the difference in performance of various classifiers is smaller than in the other experiments. Ensemble approaches like AdaBoost and Decision Tree bagged balanced with 100 estimators continue to demonstrate robust performance, reinforcing the importance of adaptive and ensemble techniques in enhancing detection sensitivity in low prevalence environments.

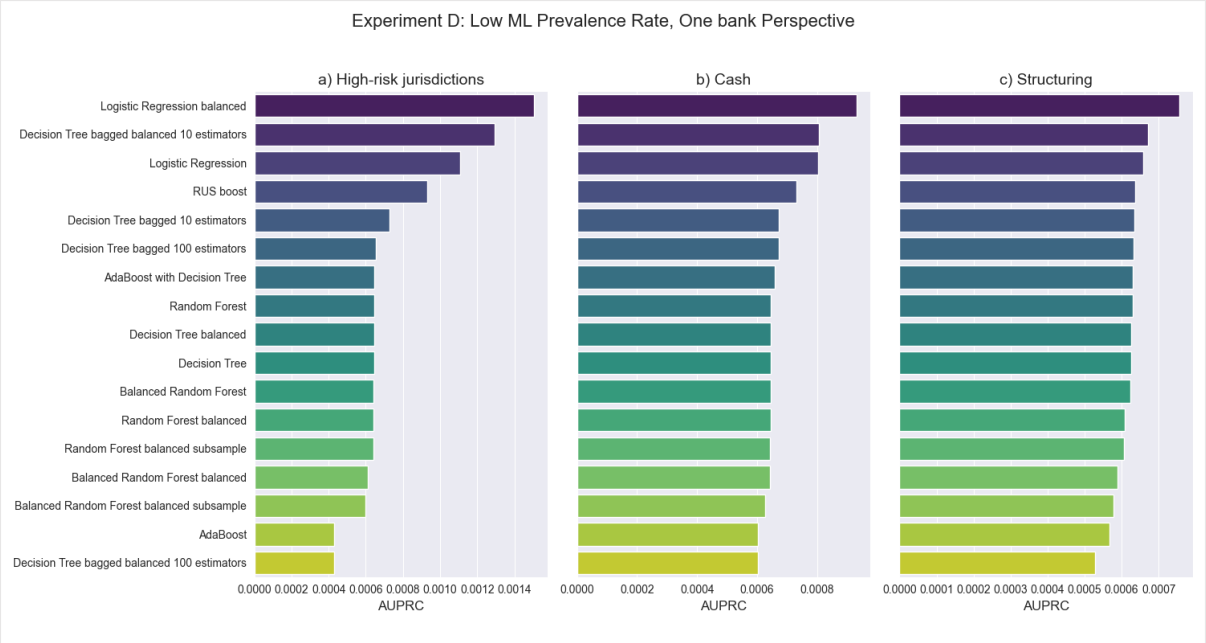

Figure 15. AUPRC scores for every model per typology in experiment D

Figure 16 evaluates the performance of different classifiers in a low ML prevalence rate environment from the perspective of a single bank. In this illustration, the Balanced Random Forest balanced classifier for high-risk jurisdictions, RUSboost for cash transactions, and Logistic Regression balanced for structuring typologies are assessed. All three models are heavily overpredicting the amount of ML transactions, exhibiting large amounts of false positives relative to true positives, indicating a strong propensity to incorrectly label non-ML transactions as ML. This indicates the poor performance of the classifiers and the lack of potential to be able to accurately estimate the FNR. This emphasizes the need for improved model specificity to accurately identify ML transactions in a low prevalence setting.

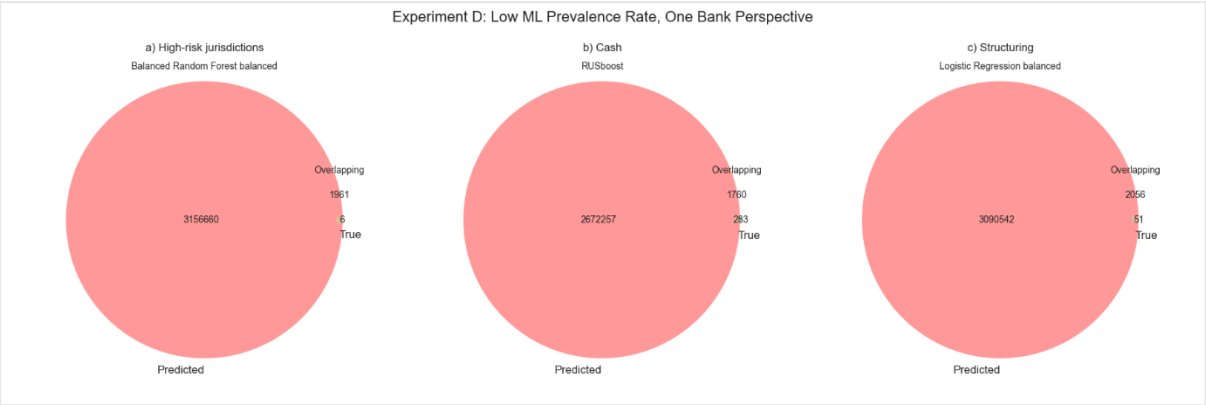

Figure 16. True and predicted number of ML transactions for the low ML rate data and the perspective of one bank

Table 18 displays a clear discrepancy between Actual and Predicted FNRs for the ML typologies. The low AUPRC values and near-zero MCC scores across all typologies reflect the model's poor precision and recall, performing marginally better than random guessing. This highlights the classifier's continued struggle in effectively identifying ML transactions.

Table 18. Results of the balanced Logistic Regression classifier for one bank on the low prevalence dataset

	High-risk jurisdictions	**Cash**	**Structuring**
Actual FNR	0.909	0.912	0.966
Predicted FNR	0.003	0.139	0.024

AUPRC	0.001	0.001	0.000
MCC	0.002	0.002	0.001

6.5 Comparison

This section provides a comparison between the average models performance over all various experiments as well as a comparison between the average model performance per typology, rate and perspective.

6.5.1 Model Comparison

This section provides an evaluation of the classification models utilized in the study, leveraging a range of metrics to ascertain their performance. The corresponding figures, which depict the average scores across these metrics over all experiments, provide a comprehensive view of model efficacy. The AUPRC, a metric tailored towards the context of imbalanced datasets, assesses the model's ability to distinguish between classes—a higher score reflects a model's improved capability to maintain a high precision rate as recall increases. Figure 17 underscores the precision-recall balance achieved by the models, with those at the top demonstrating a superior handle on the minority class prediction, which is crucial for ML classification as its prevalence is extremely limited. From this figure, we clearly have four best performing models. As the balanced bagged decision tree classifier works very similar to a random forest, we can conclude that random forest approaches with balancing measures to address the dataset imbalance are most suitable for classifying potential ML transactions. As the scale of the AUPRC is from 0 to 1, all models score relatively low. However, as the average prevalence rate of ML occurring is only 0.001 in the datasets, the score of every classifier already suggests a significant improvement in performance in comparison with random guessing or a naïve model that predicts the majority class for all instances.

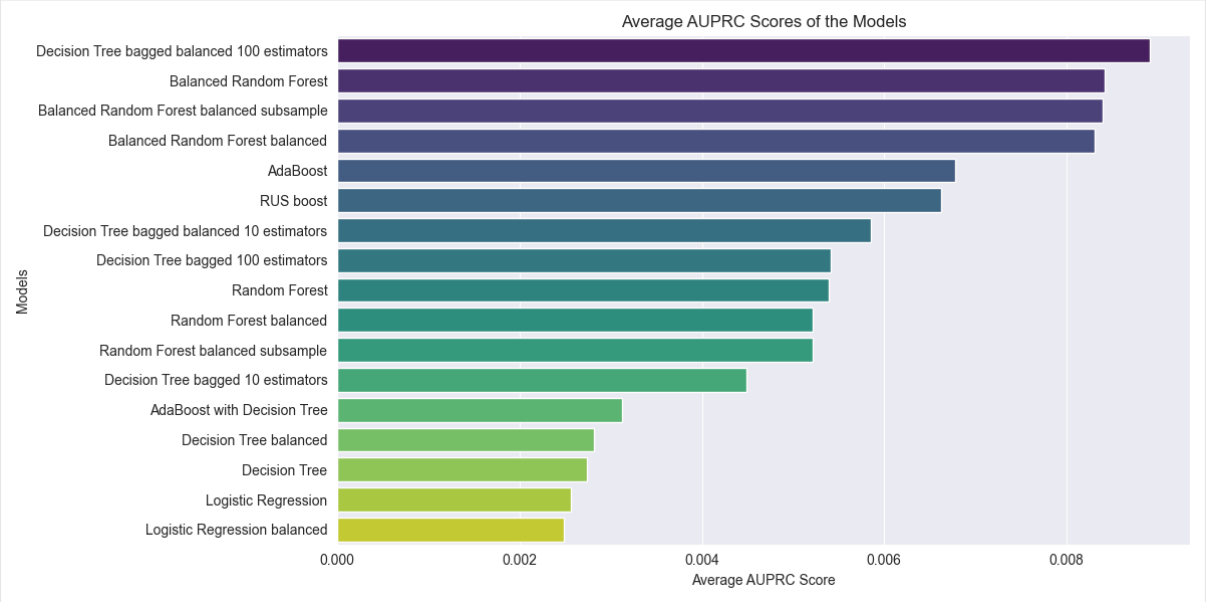

Figure 17. Average AUPRC score per model

From Figure 18, we can see that the best performing models according to this metric are different than the ranking established above. This is due to the MCC considering all four quadrants of the confusion matrix (true positives, false negatives, true negatives, and false positives) with the same priority. Due to the high class imbalance and the MCC favoring models performing well in both the majority and minority class, the majority true negatives dominate the metric due to their substantially large number. Notice how there is a certain 'leading' group with very similar performance, which drops below the RUS boost model.

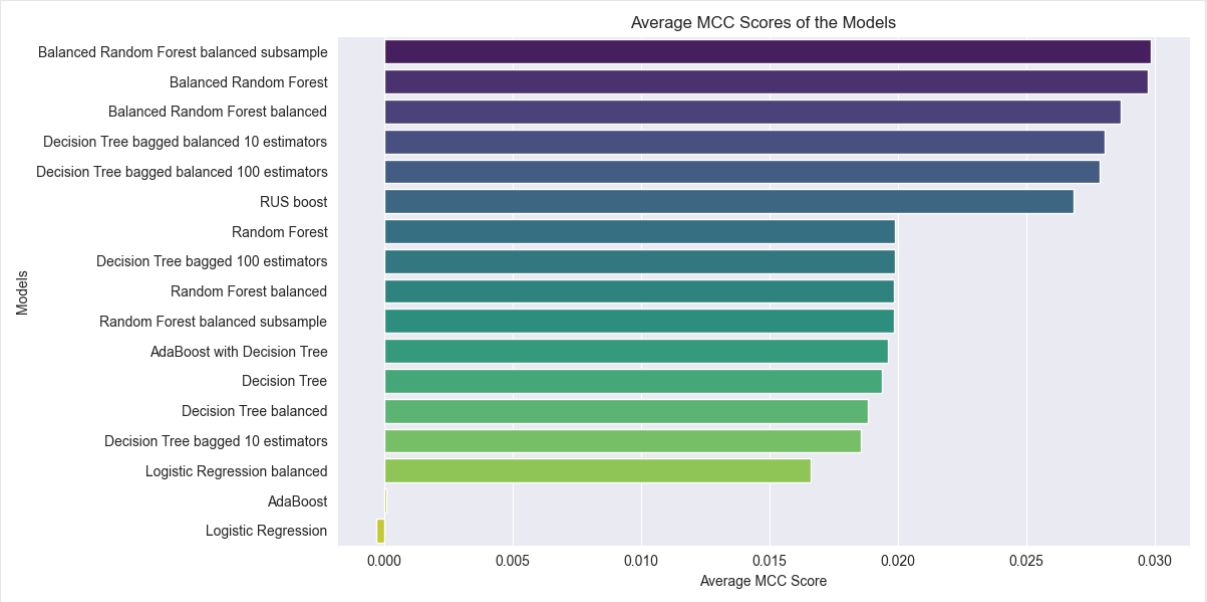

Figure 18. Average MCC score per model

6.5.2 Experiment Comparison

In this section, we discuss model performance across the three typologies, the two ML rates and the two perspectives, as depicted in Figure 19. The bar charts illustrate the average AUPRC scores. First, we can see that model performance for the cash typology is lacking in comparison to the other two typologies, especially for the high rate. This might be caused by the incredible low ML rates in the cash training dataset, as indicated in 5.2.1 Cash Typology. This indicates that if the dataset may contain too few ML transactions, classifiers will have trouble being able to learn what differentiates them from 'normal' transactions. What also is interesting to note is that the performance for both rates is higher for the high-risk jurisdictions typology than for the structuring typology, even though the ML prevalence rate of the first is with 0.22% not even half of the 0.46% of the latter. This indicates that the classifiers might have more trouble identifying certain typologies over others.

With regards to the performance across the different rates, the 'cash' typology, in particular, shows similar performance across the 'low' and 'high' rates, with AUPRC scores remaining comparably low. However, as discussed above, this behavior is expected as the ML rate in both the low and high ML rate training set has the same (low) value. On the other hand, the high-risk jurisdictions and structuring typologies show a marked performance variance between the two ML rates. The substantial difference in AUPRC scores for all three typologies indicate a sensitivity to illicit transaction volume, where higher rates potentially amplify the detection capabilities. These insights are instrumental for understanding the dynamics of model performance and the influence of transaction rates on model performance.

We conclude by comparing the two perspectives across the typologies and rates. We can clearly see that for the high-risk jurisdiction and structuring typologies, the performance differs tremendously across the various perspectives. This would indicate that the potential of combining information across banks is very large with regards to improved model performance.

For the cash typology, the perspective does not change model performance, but this once again might be because of already poor performance.

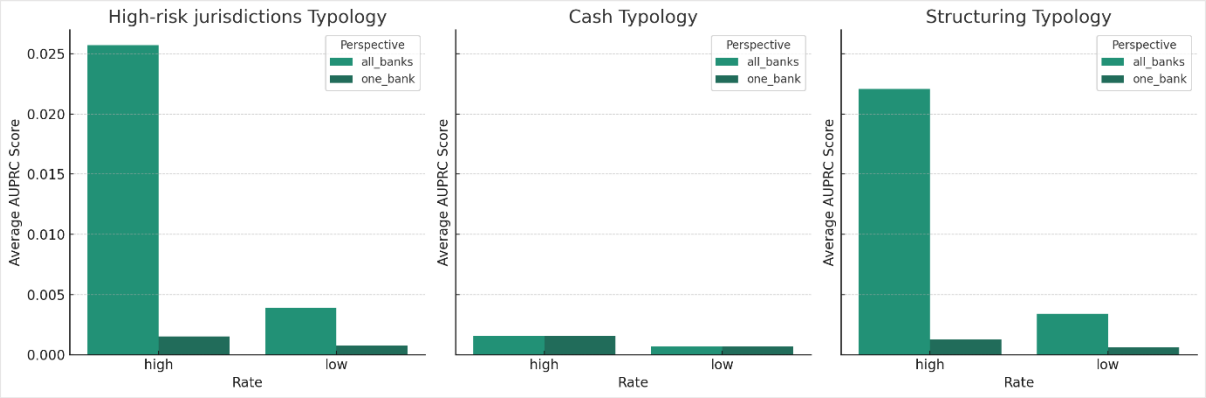

Figure 19. Average AUPRC scores by typology, rate and perspective

In all four experiments, the data suggest that balancing the class distribution, either through model adjustments or data resampling, significantly impacts the AUPRC in money laundering detection tasks. This highlights the need for specialized approaches in handling class imbalances inherent in ML datasets. Furthermore, the variability in model performance between collective and individual bank perspectives underscores the importance of context-specific model selection and training for effective ML detection.

6.5.3 Feature Importance

Finally, we provide insights into the workings of the models using Feature Importance. To be able to combine the feature importance of every type of base classifier, Permutation Feature Importance was used (Brownlee, 2020c; Scikit-learn developers, n.d.-a). This is a method to gauge the significance of features in a predictive model. It starts with training the model on the original dataset and noting its performance. The process involves shuffling the values of each feature, disrupting its relationship with the target. The model is then evaluated on this modified dataset. The degree to which the model's performance deteriorates indicates the importance of the shuffled feature. This procedure is repeated multiple times to average out random variations, providing a more reliable measure of feature significance. This technique is versatile, applicable across various models and captures feature interactions. We used the final hyperparameter tuned models and the validation dataset for evaluation.

Figure 20 presents the Aggregated Feature Importance. That means that this represents the feature importance of all models and all categories combined into one. Note that the Receiving Currency feature was removed due to having the same feature importance as the Payment Currency feature. The Payment Format feature clearly shows the highest positive feature importance, indicating a substantial and direct impact on the model's predictive capabilities. Conversely, Amount demonstrates a negative importance, suggesting an inverse relationship with the likelihood of ML. Meaning that higher transaction sums are less suspicious in the context of ML, which is in line with the insights gathered from the interviews. Lastly, Payment Currency exhibits the least importance, with a slight negative impact on the model's predictions. Collectively, these importance values suggest that the model places significant weight on the payment format when discerning ML patterns, with the amount and currency of transactions being less indicative of potential ML activities as per the synthetic datasets.

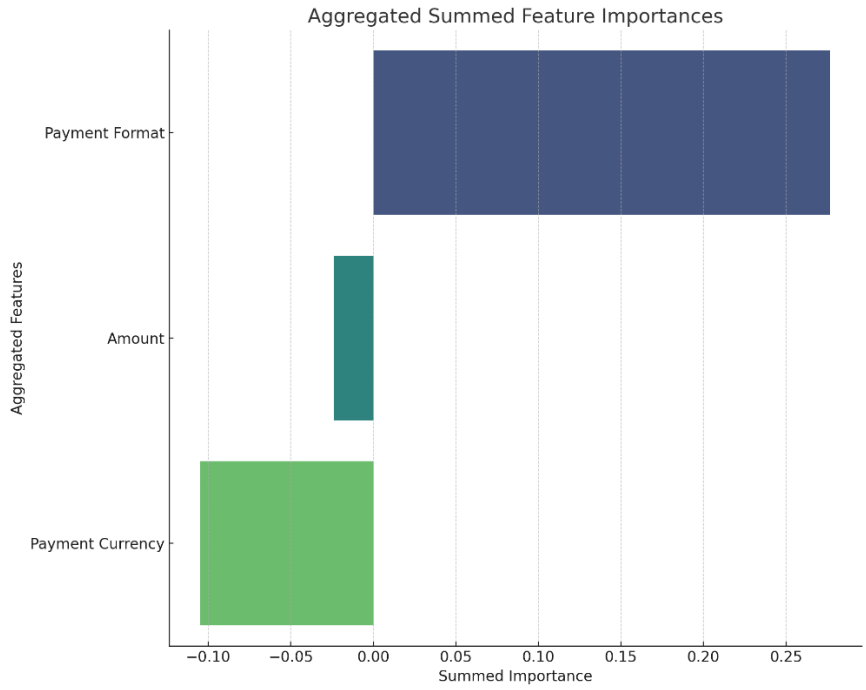

Figure 20. Aggregated Feature Importance

Figure 21 presents the feature importance of the different features individually, namely the one-hot encoded features Payment Format and Currency, as well as the continuous Amount Received variable. The features are ranked by their level of positive importance, indicated by the length and direction of the bars. The 'Payment Format – Cheque' stands out with the highest positive importance, suggesting it is a significant predictor of ML. Other payment formats like credit card, cash, and wire transfers also show a great positive feature importance, albeit to a lesser extent, pointing to their lesser but still relevant influence on the model's output. In addition, several Payment Currency categories show positive importance, such as the Euro, Shekel, and UK Pound, among others. This implies that transactions in these currencies are more likely to be flagged by the model as potential ML activities. Surprisingly, the Payment Currency Bitcoin has, although small, negative feature importance, which does not stroke with its image.

The Amount Received feature shows a very slight negative importance, indicating that it has a minimal inverse relationship with the model's predictions, indicating that lower transaction values are more likely to be flagged by the model. At the bottom, the US Dollar Payment Currency has the biggest negative importance, signifying its large 'non-ML' impact on the model's predictive behavior. Overall, the chart suggests that the payment format, particularly cheque and credit card transactions, is a key factor for the model, while the role of transaction amount and the specific currencies involved are less critical and can either slightly increase or decrease the likelihood of a transaction being flagged by the model.

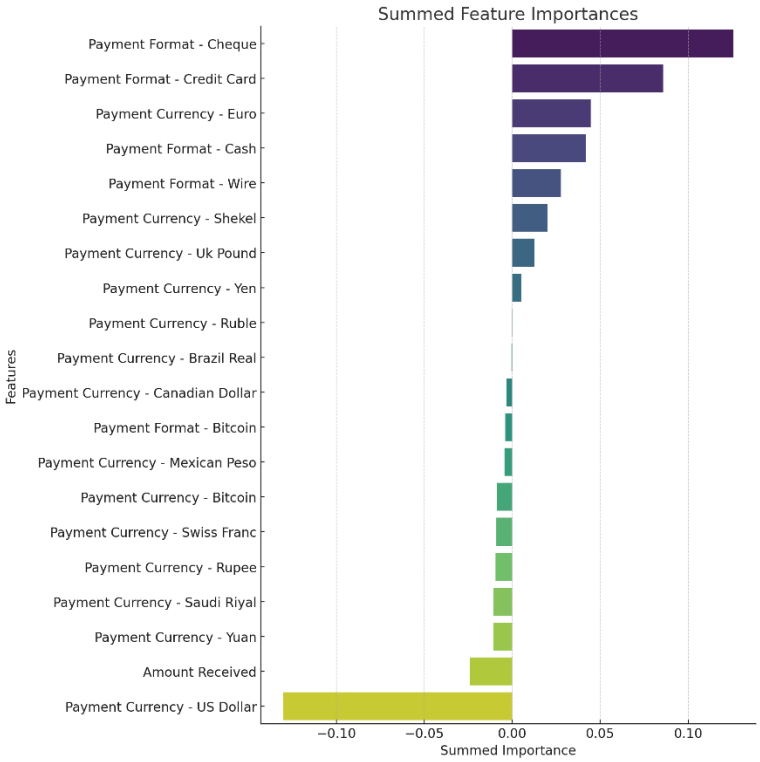

Figure 21. Feature Importance per category

7. Discussion

This chapter presents the implications, limitations and recommendations for further research in the context of estimating the FNR of rule-based TM systems within the Dutch banking sector. This study's findings have notable implications for both banks and regulators focused on identifying and mitigating transactional ML activities. As we have seen in the Results chapter, the evaluated classifiers, particularly the Balanced Random Forest variants with balancing measures, demonstrate a significant advancement over naive majority class prediction. This improvement is crucial given the extremely low prevalence rate of ML transactions in the datasets used. However, it is important to recognize that these classifiers, while beneficial in enhancing detection capabilities, show limitations in accurately estimating the FNR. The relatively modest AUPRC and MCC scores indicate challenges in predicting ML activities well. Furthermore, the difference between the actual FNR and predicted FNR is for almost all experiments and typologies very substantial, reinforcing the notion that relying on these classifiers for precise FNR estimation may be overoptimistic.

The study underscores the significance of certain indicators in flagging potential ML activities, including transaction type, payment amount, frequency, volatility, international exposure, and connections with high-risk jurisdictions. Next to those rule-based approaches, it is imperative to perceive machine learning classifiers as integral elements of a broader strategy, rather than as isolated tools for estimating the FNR. A notable observation is the variable performance of models across different ML typologies, with the structuring/smurfing and high-risk jurisdictions typologies demonstrating more distinct patterns that classifiers can learn from. This variability in performance indicates the potential advantages of adopting typology-specific approaches in TM systems, wherein models are tailored to each ML typology, considering its distinct characteristics. Additionally, the study reveals the potential benefits of collaborative approaches across banks. The marked performance differences across various perspectives, especially in the high-risk jurisdictions and structuring typologies, suggest that sharing information across institutions could significantly enhance the effectiveness of ML detection systems.

In light of existing studies, particularly those focusing on pioneering machine learning algorithms for AML purposes, this research offers novel insights. Unlike previous works that predominantly emphasize enhancing TM systems through the innovation of machine learning detection algorithms, our study pivots towards understanding the efficacy of rule-based systems and their FNR. For instance, while the foundational research by Weber et al. (2018) and subsequent developments primarily focused on the creation of more sophisticated detection algorithms, our research offers a unique perspective on the performance evaluation of existing rule-based systems. This contrast presents a significant addition to academia, highlighting areas that have received less attention in earlier studies.

This study's focus on estimating the FNR of rule-based TM systems marks a clear divergence from the dominant trend in AML research. We set out the need for a more thorough understanding of rule-based systems, which are widely utilized yet not extensively researched in existing literature. This approach underscores the potential and necessity for machine learning development to increasingly focus on models adept at handling heavy imbalanced class transaction datasets. The challenges met, such as class imbalance and the complexities in accurately estimating FNR, reinforce the necessity for future machine learning developments to prioritize these aspects. This shift in focus could significantly enhance the efficacy and deployment of machine learning TM systems in practice, something that has been lacking until now.

For regulators, the study's findings emphasize the importance of encouraging and facilitating information sharing among banks and providing guidelines for effective TM system development. Regulators should also consider the study's insights when updating compliance requirements and ML detection standards. Overall, this research offers valuable insights for stakeholders in the banking sector, providing a clearer understanding of the efficacy of various classifiers and indicators in predicting ML activities and estimating the FNR. It lays the groundwork for more sophisticated, data-driven approaches to rule-based TM evaluations, ultimately contributing to the integrity and security of the financial system. However, the study also indicates that reliance solely on these classifiers for estimating the FNR may be misguided, and a broader approach, encompassing various strategies and collaborative efforts, is essential for more effective evaluations of rule-based TM systems.

7.1 Limitations

While the study provides valuable insights, it has several limitations that must be considered. We will first discuss the quantitative limitations, after which the qualitative limitations will be presented.

The first limitation has to do with the complexity of evaluating classifier performance, as the performance of classifiers is notably intertwined with various parameters. An additional concern arises from the observation that for some experiments/typology combinations, rule-based TM systems already detect almost nothing. This indicates that the 'actual' correct FNR might be unrealistically high, casting doubt on the applicability of these findings for real-world scenarios. Such a high FNR might not accurately represent the subtleties and complexities of actual transactional behaviors, potentially leading to overestimated risks or missed opportunities for detection. The classifier parameters range from the metrics that define the performance to the parameters of the models itself which determine the trade-off between false positives and false negatives. Therefore, as the FNR is a metric of a certain model instance predicting for a certain dataset, estimating it is not straightforward and there is a risk of local optima without recognizing further potential improvements. In addition, data splits and model training processes were only executed once with constantly maintaining the same set random seed. Further research could improve this by iterating steps with different random seeds while averaging the performance. All of the above are important to consider ensuring that the model is robust and generalizable beyond just the dataset or scenarios it was trained on.

Other limitations arise from constraints of the synthetic datasets. Derived from the IT-AML simulator, they provided only eight ML patterns. Consequently, not all typologies mentioned by the interviewees were available in those datasets. While the common typologies where this study assessed for could all be incorporated somehow in the datasets, this lack of comprehensive one-on-one typological mapping could potentially impact the generalizability of the findings. In addition, as data was available for a 97 days timespan only, the potential for use of elaborate historical patterns was limited. This limitation is further exacerbated by the lack of comprehensive customer information in the datasets. This omission is significant as real-world banks do possess this information and can therefore employ Expected Transaction Profiles (ETPs). The inability to replicate such ETPs might have influenced the accuracy and relevance of the study's outcomes. Another data related limitation is caused by the removal of high-value transactions from the data. While such transactions are not typically indicative of ML, their exclusion could have led to the underrepresentation of certain ML patterns involving large sums, impacting the generalizability of our findings.

The study's findings' applicability to the Dutch banking landscape is a significant concern, particularly because the simulators and datasets were validated primarily against U.S. data, which might not accurately reflect the Dutch context. In addition, computational and time constraints limited the exploration of more advanced machine or deep learning classifiers, such as graph or neural network models. This limitation might have curtailed the accuracy and optimization of our

machine learning-based analyses. Moreover, because of computational constraints, we had to randomly sample 20% of the synthetic datasets before fitting and predicting with every model. This raises concerns about dataset representativeness, potentially omitting up to 80% of data per user. Such sampling may affect the model's generalizability across the full dataset.

The study's qualitative aspects also present limitations. The selection of experts predominantly from banking and regulatory backgrounds might have introduced bias, lacking perspectives from other relevant sectors like law enforcement or international finance and skewing the insights towards certain viewpoints. Additionally, the focus of the interviews on specific ML typologies and monitoring systems might have overlooked emerging trends and technological advancements in ML, limiting the study's scope in addressing evolving challenges in this field. Lastly, the subjective nature of qualitative data interpretation presents its own set of challenges, with the potential for unintentional biases or misinterpretations influencing the study's conclusions.

These interconnected limitations highlight the necessity for a nuanced, context-sensitive interpretation of our findings, especially in the dynamic and multifaceted arena of AML efforts within the Dutch banking sector.

7.2 Further Research

In the discussion of our research on FNR estimation using machine learning, several compelling paths for further investigation have emerged. These suggestions aim to build upon our current approach and enhance the efficacy and practicality of our models. One significant improvement lies in the development of a more balanced dataset with a substantially higher prevalence rate of ML transactions. This adjustment would enable trained models to better grasp the distinctive characteristics of such transactions. Moreover, reusing models pretrained and higher ML rate datasets can also be beneficial for performance on low ML rate datasets. Additionally, the incorporation of more advanced sampling techniques, such as oversampling the minority class or creating synthetic samples, could address class imbalance issues better and contribute to improved model generalization.

Recognizing the intricate and interconnected nature of financial transactions, another promising direction is to explore relatively new advanced graph-based machine learning methods. These techniques have the potential to capture complex relationships and patterns within financial networks that traditional models might overlook. By doing so, we can gain a deeper understanding of how funds flow through the financial system, ultimately enhancing the accuracy of ML detection. It would be beneficial to benchmark our model performance with current industry standard models.

Based on the conclusion that all banks' perspective models perform better, a pivotal area for future research involves working on data sharing while preserving privacy concerns. Collaborative efforts between banks could lead to the creation of more robust and comprehensive models, leveraging shared insights while adhering to privacy and data protection regulations. This collaboration could be crucial in developing a more effective and holistic approach to ML detection. Federated learning could be a suitable approach for this issue. To address the limitations arising from the exclusion of top 10% transaction amounts, future research should explore the impact of including these transactions on ML detection performance. Investigating the role of high-value transactions across various ML typologies can provide deeper insights and offer a more balanced representation of transactional data in ML detection systems.

Furthermore, our research can be extended by incorporating information about fund recipients and the broader financial network surrounding each account and party, as proposed by studies like Savage et al. (2016) and Colladon and Remondi (2017). While these data were not available for our current study, future research should explore additional data sources and methodologies to include this network perspective, potentially improving our ability to detect complex ML

schemes. Finally, the integration of external data sources, such as open-source financial data, economic indicators, or geopolitical events, can further enrich our analysis and enhance model performance. These additional data sources provide a broader context for transaction monitoring, potentially increasing the predictive capabilities of our models. Similarly, data spanning a longer timespan can help to improve model performance and provide better test sets for validation, potentially also helping to address evolving ML tactics over time. Future research should, at least, have greater computational resources, aiming to prevent having to implement downsampling because of computational constraints, and otherwise downsample using more representative sampling methods.

7.3 Recommendations

Following this study's findings, we can make a number of recommendations to the Dutch AML landscape.

Typology-Specific Models

Financial institutions stand at the forefront of this effort. The key is not just to adopt advanced TM systems but to tailor these models to specific ML typologies. This approach involves developing customized TM systems that are tuned for identifying the distinct patterns and behaviors characteristic of different ML activities, such as structuring/smurfing or transactions involving high-risk jurisdictions. This specialization enables a more accurate identification of suspicious activities, effectively narrowing down the focus to the most relevant indicators of ML.

Collaboration

A clear potential is the collaboration. Banks and regulators are encouraged to step, when possible, out the organizational boundaries and engage in proactive information sharing. This collective approach, underpinned by secure and compliant data exchange protocols, leverages the cumulative insights and data from multiple institutions, significantly amplifying the effectiveness of ML detection. Such collaboration not only shares the burden of detection but also enriches the pool of data, leading to more robust and comprehensive models.

Regulatory Bodies

Regulatory bodies play a dual role: firstly, by fostering an environment conducive to information sharing, and secondly, by revising ML detection standards and compliance requirements. This dual role positions regulators as both facilitators and drivers of change, pushing the sector towards more sophisticated, data-driven approaches in TM.

Evolving TM System

For TM system developers, the recommendations culminate in a call for comprehensive and flexible systems. These systems should not only incorporate a wide array of data, ranging from transaction types to customer profiles, but also be scalable and adaptable to evolving ML tactics.

8. Conclusion

In a context where Dutch banks are required by the Anti-Money Laundering and Anti-Terrorist Financing Act to implement AML measures, their task is to identify transactions that may suggest ML. However, as the amount of transactions surpassing bank systems on a daily basis is extremely large, banks are only able to manually investigate a very limited number of transactions. Therefore, their information position on what potential ML transactions they might have missed is very limited. Currently, predominantly rule-based TM systems are used to create alerts for unusual transactions. These systems, while effective in flagging transactions that cross predefined thresholds, leave a substantial subset of transactions unreviewed, thereby creating a potential risk. The question arises whether we could estimate the number of transactions that are wrongfully not flagged and therefore go unreviewed. As currently no literature exists that tries to do so, this study aimed to bridge this gap by exploring the feasibility of estimating the FNR of rule-based TM systems through supervised machine learning classifiers trained on historical alerts. Seven interviews were held to gather insights from domain experts, after which synthetic transaction datasets with labelled ML behavior were used to conduct quantitative analysis. The insights both advance the academic domain and enhance operational efficiencies within the Dutch banking sector. Prior to answering the main research question, the following four sub-questions are first addressed.

SQ1. What are the most common transactional money laundering typologies within the Dutch banking system?

The study identifies Cash, Structuring/Smurfing, and High-Risk Jurisdictions as the most prevalent transactional ML typologies in the Dutch banking system. Each of these typologies presents unique challenges for financial institutions aiming to effectively monitor and prevent ML activities. First, the Cash typology often involves unusually high deposits or withdrawals, executed in close temporal proximity. It is commonly linked to various forms of criminality. Second, the structuring/smurfing typology involves breaking down large sums of money into smaller transactions to evade mandatory reporting thresholds. A typical manifestation of this typology includes the use of multiple individuals (money mules) to deposit significant cash flows from illicit activities, such as drug trafficking, in small amounts to avoid arousing suspicion. Third, the High-Risk Jurisdictions typology pertains to transactions involving countries that are considered high-risk due to their association with ML, tax evasion, or other illicit financial activities. These transactions are subject to heightened scrutiny.

SQ2. Which indicators are most predictive of common transactional money laundering typologies in the Dutch banking sector?

The most predictive indicators for flagging potential ML activities across common typologies in the Dutch banking sector, as identified from the interviews and analysis in this study, are as follows. For the cash typology, two indicators were identified. The first one is the type of transaction. This involves differentiating between various transaction payment formats like bank transfers, cash deposits and others. The nature of the transaction can be a strong indicator of potential ML activities. The second indicator is the payment amount. This indicator is related to the size of the transaction and its comparison to static or dynamic thresholds for an account. Significant transactions, particularly those that are unusually large for a certain account type, can be indicative of ML. Second, for the structuring/smurfing typology, the first indicator is again the payment amount, similar to the cash typology, the amount of the transaction is a critical indicator, especially when small amounts are used repeatedly to avoid detection. The second indicator is the frequency of similar transactions: This considers how often similar transactions occur. A high frequency of similar transactions can be indicative of structuring, where multiple small transactions are made to evade detection thresholds. Third, the volatility

in transactions: rapid and significant fluctuations in account cashflows can indicate structuring activities, as funds are moved in and out of an account quickly. Finally, for the high-risk jurisdictions typology, the first indicator is the involvement of high-risk international jurisdictions: transactions that involve countries known for higher ML risks are particularly scrutinized. Next to that, a large proportion of transactions involving foreign entities can be an indicator, especially when this is high or unusual for the account in question. Finally, again the transaction amount involved is a key indicator, especially in relation to international transactions involving high-risk jurisdictions.

SQ3. What are the thresholds used for key indicators in the rule-based transaction monitoring systems employed by Dutch banks?

Dutch banks use both static and dynamic thresholds in their rule-based TM systems, although dynamic indicators are generally considered more effective than static ones in capturing the evolving complexities of ML activities. The use of both static and dynamic thresholds within rule-based TM systems allows for a nuanced approach to detecting potential ML activities, balancing the need for comprehensive monitoring with the practicalities of handling large volumes of transactional data. These thresholds are tailored according to the risk category of a client or product, as well as the specific typology under scrutiny. For instance:

- For the Cash typology, alerts could be generated for "cash deposits above amount X."
- For the Structuring/Smurfing typology, alerts could be generated for "more than X% of your average account amount."
- For the High-Risk Jurisdictions typology, heightened scrutiny is applied to transactions involving or originating from countries considered high-risk.

The application of thresholds is complex and subject to continual tuning to maintain a balance between identifying true positives and minimizing false positives. Importantly, the thresholds are influenced by the Expected Transaction Profile (ETP) of clients and can vary significantly depending on the market in which a bank operates, the complexity of its products, and the characteristics of its client population.

SQ4. Which supervised machine learning classifiers, recommended by literature, demonstrate the highest efficacy in predicting unusual transactions indicative of money laundering?

As advancements in the field follow each other up rather quickly, many different machine learning classifiers are used within literature. However, the most fundamental supervised machine learning classifiers for predicting unusual transactions indicative of ML are found to be Logistic Regression, Decision Trees, and Random Forests, along with their variants that address class imbalance. These simple and computationally cheap classifiers often still demonstrate remarkably high performance, especially considering their efficiency. Logistic Regression excels in interpreting and predicting binary outcomes. Decision Trees are advantageous for their ability to handle diverse data types without extensive preprocessing and for capturing complex patterns in data, crucial for identifying unusual ML transactions. Random Forests and their ensemble counterparts, such as Balanced Random Forest and AdaBoost with decision trees, demonstrate superior performance. These models effectively handle imbalanced datasets, a common challenge in detecting ML activities. They enhance performance by incorporating downsampling and cost-sensitive learning, focusing on the minority class and improving false negative detection.

The main research question at hand is:

To what extent could a supervised machine learning classifier, when trained on historical alerts, assist Dutch banks in estimating the False Negative Rate of rule-based transaction monitoring systems concerning unreviewed transactions?

In addressing the main research question, the study demonstrates that supervised machine learning classifiers cannot assist Dutch banks well in estimating the FNR of rule-based transaction

monitoring systems for unreviewed transactions. The study's findings reveal a pronounced discrepancy between the actual and predicted FNRs across different experiments and typologies, underscoring the classifiers' generally poor performance. Notably, in high ML prevalence settings, even the better-performing models, such as the balanced bagging classifier or ensemble models with balancing measures, showed substantial gaps between actual FNR (ranging from 0.731 to 0.990 across typologies) and predicted FNR (ranging from 0.003 to 0.986). This was further compounded in low prevalence scenarios and single bank perspectives, where classifiers predominantly over-predicted non-ML transactions, leading to high false positives and minimal detection of actual ML activities. The low AUPRC and MCC scores across various contexts highlight the significant challenges in using machine learning for effective ML transaction detection in Dutch banking.

References

Abegaz, T., Berhane, Y., Worku, A., Assrat, A., & Assefa, A. (2014). Road Traffic Deaths and Injuries Are Under-Reported in Ethiopia: A Capture-Recapture Method. *PLOS ONE*, *9*(7), e103001. https://doi.org/10.1371/journal.pone.0103001

Adams, W. C. (2015). Conducting Semi-Structured Interviews. In *Handbook of Practical Program Evaluation* (pp. 492–505). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119171386.ch19

Alkhalili, M., Qutqut, M. H., & Almasalha, F. (2021). Investigation of Applying Machine Learning for Watch-List Filtering in Anti-Money Laundering. *IEEE Access*, *9*, 18481–18496. https://doi.org/10.1109/ACCESS.2021.3052313

Altman, E. (2023). *IBM Transactions for Anti Money Laundering (AML)* [Data set]. https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml

Altman, E., Egressy, B., Blanuša, J., & Atasu, K. (2023). *Realistic Synthetic Financial Transactions for Anti-Money Laundering Models* (arXiv:2306.16424). arXiv. https://doi.org/10.48550/arXiv.2306.16424

AMLC. (2020, July 28). *Money laundering indicators*. AMLC.EU. https://www.amlc.eu/money-laundering-indicators/

Baazil, D. (2022, April 20). De bijl aan de brievenbus. *De Groene Amsterdammer*, *16*. https://www.groene.nl/artikel/de-bijl-aan-de-brievenbus

Bank for International Settlements. (2023). *Project Aurora: The power of data, technology and collaboration to combat money laundering across institutions and borders*. https://www.bis.org/publ/othp66.htm

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Brittain, S., & Böhning, D. (2009). Estimators in capture–recapture studies with two sources. *AStA Advances in Statistical Analysis*, *93*(1), 23–47. https://doi.org/10.1007/s10182-008-0085-y

BroniD. (2019, October 28). *What are the Typologies of Money Laundering and Terrorist Financing?* https://www.bronid.com/post/what-are-the-typologies-of-money-laundering-and-terrorist-financing

Brownlee, J. (2020a, January 12). How to Fix k-Fold Cross-Validation for Imbalanced Classification. *MachineLearningMastery.Com*. https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/

Brownlee, J. (2020b, February 9). A Gentle Introduction to Threshold-Moving for Imbalanced Classification. *MachineLearningMastery.Com*. https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

Brownlee, J. (2020c, March 29). How to Calculate Feature Importance With Python. *MachineLearningMastery.Com*. https://machinelearningmastery.com/calculate-feature-importance-with-python/

Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, *131*, 441–452. https://doi.org/10.1016/j.jbusres.2020.10.012

Cassara, J. A. (2015). *Trade-Based Money Laundering: The Next Frontier in International Money Laundering Enforcement*. John Wiley & Sons.

Chau, D., & Nemcsik, M. van D. (2020). *Anti-Money Laundering Transaction Monitoring Systems Implementation: Finding Anomalies*. John Wiley & Sons.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.

Confusion matrix. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=1148699071

Connors, B. M., Cooper, A. B., Peterman, R. M., & Dulvy, N. K. (2014). The false classification of extinction risk in noisy environments. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1787), 20132935. https://doi.org/10.1098/rspb.2013.2935

Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

DNB. (2020). *Leidraad Wwft en Sanctiewet*. https://www.dnb.nl/voor-de-sector/open-boek-toezicht/wet-regelgeving/wwft/dnb-leidraad-wwft-en-sw/

DNB. (2021, January 26). *Introduction Wwft*. DeNederlandscheBank. https://www.dnb.nl/en/sector-information/supervision-laws-and-regulations/laws-and-eu-regulations/anti-money-laundering-and-anti-terrorist-financing-act/introduction-wwft/

Eddin, A. N., Bono, J., Aparício, D., Polido, D., Ascensão, J. T., Bizarro, P., & Ribeiro, P. (2022). *Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs* (arXiv:2112.07508). arXiv. https://doi.org/10.48550/arXiv.2112.07508

Elliptic. (2020, December 10). *What Are Money Laundering Typologies?* https://www.elliptic.co/blog/what-are-typologies-in-money-laundering

European Banking Authority. (2018). *Guidelines on risk factors and simplified and enhanced customer due diligence*. https://www.eba.europa.eu/regulation-and-policy/anti-money-laundering-and-e-money/guidelines-on-risk-factors-and-simplified-and-enhanced-customer-due-diligence

European Commission. (2021, July 20). *Beating financial crime: Commission overhauls anti-money laundering and countering the financing of terrorism rules* [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3690

EY. (2018). *Anti-money laundering (AML) Transaction Monitoring*. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/emeia-financial-services/ey-anti-money-laundering-aml-transaction-monitoring.pdf

FATF. (2006). *Trade-Based Money Laundering*. https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Methodsandtrends/Trade-basedmoneylaundering.html

FATF – Egmont Group. (2020). *Trade-Based Money Laundering: Risk Indicators*. FATF. https://www.fatf-gafi.org/en/publications/Methodsandtrends/Trade-based-money-laundering-indicators.html

FIAU. (2021). *Typologies & Red Flags: Indicators of Tax-Related ML*. https://fiaumalta.org/wp-content/uploads/2021/11/FIAU-Factsheet-Typologies-Red-Flags-Indicators-of-Tax-Related-ML.pdf

FinCEN. (n.d.). *What is money laundering?* Financial Crimes Enforcement Network. https://www.fincen.gov/what-money-laundering

FinCEN. (2010). *Advisory to Financial Institutions on Filing Suspicious Activity Reports regarding Trade-Based Money Laundering*. https://www.fincen.gov/resources/advisories/fincen-advisory-fin-2010-a001

FIU-Nederland. (n.d.-a). *Kennisbank*. FIU-Nederland.

FIU-Nederland. (n.d.-b). *Wwft (Prevention) Act*. Financial Intelligence Unit - the Netherlands. https://www.fiu-nederland.nl/en/legislation/general-legislation/wwft

FIU-Nederland. (2021, March 18). *Veelgestelde vragen*. FIU-Nederland. https://www.fiu-nederland.nl/home/veelgestelde-vragen/

FIU-Nederland. (2022). *Jaaroverzicht 2021*. https://www.fiu-nederland.nl/home/jaaroverzichten/

Fronzetti Colladon, A., & Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications*, *67*, 49–58. https://doi.org/10.1016/j.eswa.2016.09.029

Gerlings, J., & Constantiou, I. (2022). *Machine Learning in Transaction Monitoring: The Prospect of xAI* (arXiv:2210.07648). arXiv. https://doi.org/10.48550/arXiv.2210.07648

Goldberg, J. D., & Wittes, J. T. (1978). The Estimation of False Negatives in Medical Screening. *Biometrics*, *34*(1), 77–86. https://doi.org/10.2307/2529590

Han, J., Huang, Y., Liu, S., & Towey, K. (2020). Artificial intelligence for anti-money laundering: A review and extension. *Digital Finance*, *2*(3), 211–239. https://doi.org/10.1007/s42521-020-00023-1

Harding, L. (2017, March 20). The Global Laundromat: How did it work and who benefited? *The Guardian*. https://www.theguardian.com/world/2017/mar/20/the-global-laundromat-how-did-it-work-and-who-benefited

Hasanin, T., & Khoshgoftaar, T. (2018). The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 70–79. https://doi.org/10.1109/IRI.2018.00018

Herzog, R., Elgort, D. R., Flanders, A. E., & Moley, P. J. (2017). Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period. *The Spine Journal*, *17*(4), 554–561. https://doi.org/10.1016/j.spinee.2016.11.009

Imbalanced-learn developers. (n.d.-a). *BalancedBaggingClassifier*. https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedBaggingClassifier.html

Imbalanced-learn developers. (n.d.-b). *BalancedRandomForestClassifier*. https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html

Imbalanced-learn developers. (n.d.-c). *Imbalanced-learn*. https://imbalanced-learn.org/stable/#

Imbalanced-learn developers. (n.d.-d). *RUSBoostClassifier*. https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.RUSBoostClassifier.html

International Centre for Asset Recovery (ICAR). (2022). *Basel AML Index*. https://index.baselgovernance.org

Interpol. (2023, May 9). *Money Laundering*. https://www.interpol.int/en/Crimes/Financial-crime/Money-laundering

Jass, KYC-Chain. (2019, April 25). *The History Of Money Laundering*. https://kyc-chain.com/the-history-of-money-laundering/, https://kyc-chain.com/the-history-of-money-laundering/

JetBrains. (n.d.). *JetBrains DataSpell: The IDE for data analysis*. JetBrains. https://www.jetbrains.com/dataspell/

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*, *24*(4), 602–611. https://doi.org/10.2307/2392366

J.P. Morgan AI Research. (n.d.). *Anti-Money Laundering (AML)*. Retrieved June 26, 2023, from https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data/anti-money-laundering

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, *23*(1), 173–186. https://doi.org/10.1108/JMLC-07-2019-0055

Kenton, W. (2022, May 21). *Anti Money Laundering (AML) Definition: Its History and How It Works*. Investopedia. https://www.investopedia.com/terms/a/aml.asp

KLPD - Dienst Nationale Recherche Informatie. (n.d.). *Indicatorenlijst behorende bij de WWFT*.

KnowYourCountry. (n.d.-a). *Brazil Country Summary*. https://www.knowyourcountry.com/brazil

KnowYourCountry. (n.d.-b). *China Country Summary*. https://www.knowyourcountry.com/china

KnowYourCountry. (n.d.-c). *Mexico Country Summary*. https://www.knowyourcountry.com/mexico

Levi, M., & Reuter, P. (2006). Money laundering. *Crime and Justice*, *34*(1), 289–375.

Lopez, E. (2023). *EdgarLopezPhD/PaySim* [Java]. https://github.com/EdgarLopezPhD/PaySim (Original work published 2017)

Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). *PaySim: A financial mobile money simulator for fraud detection*. 249–255.

Mane, S., Srivastava, J., Hwang, S.-Y., & Vayghan, J. (2004). Estimation of false negatives in classification. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 475–478. https://doi.org/10.1109/ICDM.2004.10048

Ministerie van Algemene Zaken. (2022, September 23). *Verdere verbetering van aanpak van witwassen—Nieuwsbericht—Rijksoverheid.nl* [Nieuwsbericht]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/actueel/nieuws/2022/09/23/verdere-verbetering-van-aanpak-van-witwassen

Uitvoeringsbesluit Wwft 2018. https://wetten.overheid.nl/BWBR0041193/2022-11-01

Wet ter voorkoming van witwassen en financieren van terrorisme. https://wetten.overheid.nl/BWBR0024282/2022-11-01

Morse, J. M. (1991). Approaches to Qualitative-Quantitative Methodological Triangulation. *Nursing Research*, *40*(2), 120.

NVB. (2023, April). *NVB Standaarden/NVB Risk Based Industry Baselines*. Nederlandse Vereniging van Banken. https://www.nvb.nl/publicaties/protocollen-regelingen-richtlijnen/nvb-standaardennvb-risk-based-industry-baselines/

Paxton, R. (2015, November 25). *From the laundromat to Wall Street: A history of money laundering* [Medium]. https://medium.com/@alacergroup/from-the-laundromat-to-wall-street-a-history-of-money-laundering-c6a5407e785c

Python Software Foundation. (2023, November 23). *Python.org*. Python.Org. https://www.python.org/

Razzak, J. A., & Luby, S. P. (1998). Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method. *International Journal of Epidemiology*, *27*(5), 866–870. https://doi.org/10.1093/ije/27.5.866

Reuter, P., & M. Truman, E. (2005). *Chasing dirty money: The fight against money laundering*. Peterson Institute.

Rocha-Salazar, J.-J., Segovia-Vargas, M.-J., & Camacho-Miñano, M.-M. (2021). Money laundering and terrorism financing detection using neural networks and an abnormality indicator. *Expert Systems with Applications*, *169*, 114470. https://doi.org/10.1016/j.eswa.2020.114470

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Savage, D., Wang, Q., Chou, P., Zhang, X., & Yu, X. (2016). *Detection of money laundering groups using supervised learning in networks* (arXiv:1608.00708). arXiv. https://doi.org/10.48550/arXiv.1608.00708

Scikit-learn developers. (n.d.-a). *4.2. Permutation feature importance*. Scikit-Learn. https://scikit-learn/stable/modules/permutation_importance.html

Scikit-learn developers. (n.d.-b). *AdaBoostClassifier*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

Scikit-learn developers. (n.d.-c). *BaggingClassifier*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

Scikit-learn developers. (n.d.-d). *Compare the effect of different scalers on data with outliers*. Scikit-Learn. https://scikit-learn/stable/auto_examples/preprocessing/plot_all_scaling.html

Scikit-learn developers. (n.d.-e). *DecisionTreeClassifier*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Scikit-learn developers. (n.d.-f). *LogisticRegression*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit-learn developers. (n.d.-g). *RandomForestClassifier*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit-learn developers. (n.d.-h). *Scikit-learn: Machine learning in Python*. https://scikit-learn.org/stable/index.html

Southwood, T. R. E., & Henderson, P. A. (2009). *Ecological Methods*. John Wiley & Sons.

Statista. (2023a, July 21). *Money laundering and terrorist financing Mexico 2022*. Statista. https://www.statista.com/statistics/877359/risk-index-money-laundering-terrorist-financing-mexico/

Statista. (2023b, October 23). *Money laundering and terrorist financing in Brazil 2021*. Statista. https://www.statista.com/statistics/877247/risk-index-money-laundering-terrorist-financing-brazil/

Sun, X., Feng, W., Liu, S., Xie, Y., Bhatia, S., Hooi, B., Wang, W., & Cheng, X. (2022). MonLAD: Money laundering agents detection in transaction streams. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 976–986.

Suzumura, T., & Kanezashi, H. (2023). *Anti-Money Laundering Datasets (AMLSim)* [Python]. International Business Machines. https://github.com/IBM/AMLSim (Original work published 2018)

Teeffelen, K. van. (2018, November 5). Nog altijd wassen criminelen in Nederland miljarden euro's wit. *Trouw*. https://www.trouw.nl/nieuws/nog-altijd-wassen-criminelen-in-nederland-miljarden-euro-s-wit~be172327/

Tertychnyi, P., Slobozhan, I., Ollikainen, M., & Dumas, M. (2020). Scalable and Imbalance-Resistant Machine Learning Models for Anti-money Laundering: A Two-Layered Approach. In B. Clapham & J.-A. Koch (Eds.), *Enterprise Applications, Markets and Services in the Finance Industry* (pp. 43–58). Springer International Publishing. https://doi.org/10.1007/978-3-030-64466-6_3

Transactie Monitoring Nederland. (n.d.). *Collaborating towards effective transaction monitoring*. TMNL. https://tmnl.nl/en/

UKALA. (2012, April 25). *A Brief History of Money Laundering*. https://www.ukala.org.uk/ukala_library_resour/a-brief-history-of-money-laundering/

Unger, B., & Van Waarden, F. (2009). How to dodge drowning in data? Rule-and risk-based anti money laundering policies compared. *Review of Law & Economics*, *5*(2), 953–985.

UNODC. (n.d.-a). *Money Laundering*. United Nations : Office on Drugs and Crime. Retrieved July 11, 2023, from //www.unodc.org/unodc/en/money-laundering/overview.html

UNODC. (n.d.-b). *SDG16: Peace and Justice*. United Nations : Office on Drugs and Crime. //www.unodc.org/unodc/en/sustainable-development-goals/sdg16_-peace-and-justice.html

UNODC. (2010). *Risk of Money Laundering through Financial Instruments* (p. 130). https://www.unodc.org/documents/colombia/2013/diciembre/Risk_of_Money_Lauder ing_version_I.pdf

Vassallo, D., Vella, V., & Ellul, J. (2021). Application of Gradient Boosting Algorithms for Anti-money Laundering in Cryptocurrencies. *SN Computer Science*, *2*(3), 143. https://doi.org/10.1007/s42979-021-00558-z

Weber, M., Chen, J., Suzumura, T., Pareja, A., Ma, T., Kanezashi, H., Kaler, T., Leiserson, C. E., & Schardl, T. B. (2018). *Scalable Graph Learning for Anti-Money Laundering: A First Look* (arXiv:1812.00076). arXiv. https://doi.org/10.48550/arXiv.1812.00076

Wheeler, J. (2021, October 28). *10 AML Rules for Your Compliance Program | Jumio*. Jumio: End-to-End ID, Identity Verification and AML Solutions. https://www.jumio.com/10-sample-aml-compliance-rules/

Zhang, Y., & Trubey, P. (2019). Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. *Computational Economics*, *54*(3), 1043–1063. https://doi.org/10.1007/s10614-018-9864-z

Appendices

Appendix A. Literature Review

Literature review searches were performed using the Google Scholar search engine and the Scopus and TU Delft WorldCat databases. The following (combinations of) search queries have been used to conduct the literature review:

Table 19. An overview of the search queries used to conduct the literature review.

Primary search term(s)	**Additional search term(s)**	**Number of results on**		
		Google Scholar	Scopus	WorldCat
Money laundering	typologies	19.000	51	177
Transaction monitoring	money laundering typologies	10.800	2	56
	anti-money laundering	14.600	53	729
	anti-money laundering + statistics	21.500	1	196
Synthetic data	money laundering	14.100	22	36
	money laundering + transaction monitoring	18.200	1	96
Machine learning	money laundering + transaction monitoring	7.420	8	31.000
	transaction monitoring	67.400	128	47.800
	money laundering	17.300	1	56.500
Statistical similarity	curve	2.430.000	1.291	100.000
	distribution	5.710.000	6.045	284.000

Appendix B. Interview Questions

1. What is your position and expertise?
2. What are the three most common transactional money laundering typologies within the Dutch banking system?
3. By which three indicators can these money laundering typologies best be recognized?
    a. Do these indicators operate individually from each other or collectively per scenario?
    b. Do these indicators operate based on probability or risk (probability * impact)?
4. What type of thresholds (static, dynamic or both) and threshold values are used for these three indicators within rule-based transaction monitoring systems?
    a. Do static and dynamic rules work "together" in the same rules or independently side by side?
    b. How many such rules exist in the average system of a Dutch bank?
    c. Do these rules normally consist of a "parameter" or are there more conditions within a rule? E.g., conditional rules?
    d. Does a rule (or a system) give a binary "yes/no" answer or a risk score?
5. How do you estimate the risk probability distributions of each money laundering typology?
6. How much information does a bank receive after making a FIU report?

Appendix C. Hyperparameter tuning

This appendix presents the parameters and ranges which have been used in the random search cross-validation.

Table 20. Hyperparameter tuning setup

Classifier	Sample & Feature Parameters
BalancedBaggingClassifier	max_samples: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0], max_features: [0.5, 0.75, 1.0], bootstrap: [True, False], max_depth: [5, 10, 20, 30, None], min_samples_split: [2, 4, 6, 8, 10], min_samples_leaf: [1, 2, 4, 6], criterion: ['gini', 'entropy']
BalancedRandomForestClassifier	max_samples: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0], max_features: ['sqrt', 'log2'], bootstrap: [True, False]
AdaBoostClassifier	max_depth: [1, 2, 5, 10, 20, 30, None], min_samples_split: [2, 4, 6, 8, 10], min_samples_leaf: [1, 2, 4, 6], criterion: ['gini', 'entropy'], learning_rate: [0.01, 0.05, 0.1, 0.2, 0.5, 1.0], algorithm: ['SAMME', 'SAMME.R'], max_features: ['sqrt', 'log2', None]
RUSBoostClassifier	learning_rate: [0.01, 0.05, 0.1, 0.2, 0.5, 1.0], algorithm: ['SAMME', 'SAMME.R'], max_features: ['sqrt', 'log2', None]
LogisticRegression	solver: ['liblinear', 'lbfgs', 'newton-cg'], penalty: ['l1', 'l2'], C: loguniform(1e-5, 100)