АВТОМАТТАНДЫРУ ЖӘНЕ БАСҚАРУ
АВТОМАТИЗАЦИЯ И УПРАВЛЕНИЕ
AUTOMATION AND CONTROL

DOI 10.51885/1561-4212_2024_2_53
IRSTI 50.43.15

A.K. Kulbayeva[1], S.B. Rakhmetulayeva[2], A.K. Bolshibayeva[3]
International Information Technology University, Almaty, Kazakhstan
*[1]E-mail: aakulbayeva@gmail.com**
[2]E-mail: ssrakhmetulayeva@gmail.com
[3]E-mail: kakim-aigerim@mail.ru

DETECTING MONEY LAUNDERING ACTIVITIES IN KAZAKHSTAN: A MACHINE LEARNING APPROACH AND A COMPREHENSIVE STUDY

ҚАЗАҚСТАНДА АҚШАНЫ ЖЫМҚЫРУ ЖӨНІНДЕГІ ҚЫЗМЕТТІ АНЫҚТАУ: МАШИНАЛЫҚ ОҚЫТУ ТӘСІЛІ ЖӘНЕ КЕШЕНДІ ЗЕРТТЕУ

ВЫЯВЛЕНИЕ ДЕЯТЕЛЬНОСТИ ПО ОТМЫВАНИЮ ДЕНЕГ В КАЗАХСТАНЕ: ПОДХОД МАШИННОГО ОБУЧЕНИЯ И КОМПЛЕКСНОЕ ИЗУЧЕНИЕ

Abstract. *In the context of the global challenge of money laundering, this study conducts a comprehensive national risk assessment, with a focus on Kazakhstan. The research employs state-of-the-art methods to identify vulnerabilities within both financial and non-financial sectors and assess the potential risks associated with money laundering. The study uses innovative methodologies, including unsupervised and supervised learning techniques, to analyze patterns in financial transactions, aiming to distinguish between legitimate operations and potential money laundering activities. The application of K-means clustering, and logistic regression reveals promising results in detecting anomalies and suspicious transactions. By incorporating synthetic financial transaction data, the research provides insights into money laundering practices and their concealed nature. This study serves as an initial step in enhancing anti-money laundering efforts and strengthening the legal and institutional framework in Kazakhstan. The findings offer valuable insights into the detection of money laundering and its implications for national and international security.*

Keywords: *Money laundering detection, National risk assessment, Anti-money laundering (AML), Terrorism financing, Machine learning, Logistic regression, K-means clustering.*

Аңдатпа. *Ақшаны жымқырудың жаһандық мәселесі шеңберінде осы зерттеуде Қазақстанға баса назар аудара отырып, ұлттық тәуекелдерді жан-жақты бағалау жүргізіледі. Зерттеу қаржылық және қаржылық емес секторлардағы осалдықтарды анықтау және ақшаны жымқыруға байланысты ықтимал тәуекелдерді бағалау үшін озық әдістерді пайдаланады. Заңды операцияларды және ақшаны жымқырудың ықтимал схемаларын ажырату мақсатында қаржылық транзакциялардың заңдылықтарын талдау үшін зерттеуде машиналық оқыту, яғни, мұғалімсіз оқыту және мұғаліммен оқыту әдістерін қоса алғанда, инновациялық әдіснамалар қолданылады. K-орташа және логистикалық регрессия алгоритмін қолдану аномалиялар мен күдікті транзакцияларды анықтауда жүйелі нәтижелер көрсетеді. Қаржылық операциялардың синтетикалық деректерін қосу зерттеуге ақшаны жымқыру тәжірибесін зерттеуге мүмкіндік береді. Бұл зерттеу ақшаны жымқыруға қарсы күрес және Қазақстандағы құқықтық және институционалдық базаны нығайту жөніндегі күш-жігерді күшейтудегі бастапқы қадам ретінде қызмет етеді. Алынған нәтижелер ақшаны жымқыруды анықтау және оның ұлттық және әлемдік қауіпсіздікке әсері туралы құнды ғылыми тұжырымдар береді.*

Түйін сөздер: *ақшаны жылыстатуды анықтау, тәуекелдерді ұлттық бағалау, ақшаны жылыстатуға қарсы іс-қимыл, терроризмді қаржыландыру, Машиналық оқыту, логистикалық регрессия, K-орташа алгоритм.*

Аннотация. *В рамках глобальной проблемы отмывания денег в данном исследовании проводится всесторонняя оценка национальных рисков с акцентом на Казахстан. В исследовании используются передовые методы для выявления уязвимостей в финансовом и нефинансовом*

секторах и оценки потенциальных рисков, связанных с отмыванием денег. Для анализа паттернов финансовых транзакций с целью различения законных операций и потенциальных схем отмывания денег в исследовании применяются инновационные методологии, включая методы обучения без учителя и обучения с учителем. Применение алгоритма K-средних и логистической регрессии показывает многообещающие результаты в выявлении аномалий и подозрительных транзакций. Включение синтетических данных о финансовых операциях позволяет исследованию раскрывать практики отмывания денег и их скрытую природу. Это исследование служит начальным шагом в усилении усилий по борьбе с отмыванием денег и укреплению правовой и институциональной базы в Казахстане. Полученные результаты предоставляют ценные научные выводы относительно выявления отмывания денег и его влияния на национальную и мировую безопасность.

Ключевые слова: *выявление отмывания денег, национальная оценка рисков, противодействие отмыванию денег (ПОД), финансирование терроризма, машинное обучение, логистическая регрессия, алгоритм K-средних.*

Introduction. Money laundering represents a problem worth billions of dollars. It's an exceptionally challenging task to detect money laundering activities. Most automated algorithms tend to produce a high number of false positives, where lawful transactions are mistakenly identified as money laundering [1]. Conversely, there is a significant concern regarding false negatives, which are instances of money laundering going undetected. Criminals, as expected, make significant efforts to hide their trail.

The national risk assessment of money laundering has the following goals:

1. Identifying commonly used money laundering schemes.
2. Identifying vulnerabilities in both the financial and non-financial sectors, as well as in existing laws.
3. Promoting a unified understanding of money laundering risks at the national level among the Financial Monitoring Committee (FMC), government entities, law enforcement agencies, and specialized organizations.
4. Developing measures to minimize and effectively manage money laundering risks.

To accomplish the objectives of this research, specific tasks for the national risk assessment have been outlined:

1. Identifying threats and vulnerabilities related to money laundering stemming from predicate and high-risk crimes.
2. Examining the enforcement practices of legislation by government bodies involved in combating money laundering and terrorist financing.
3. Analyzing the criminogenic environment to pinpoint the factors and circumstances that facilitate money laundering.
4. Formulating comprehensive anti-money laundering and counter-terrorist financing strategies within the Republic of Kazakhstan.

Known cases of money laundering and terrorist financing can be sensitive information and their details are often not published publicly. However, in the past there have been some measures related to money laundering and terrorist financing in the Republic of Kazakhstan [2].

1. Case of financial support in the West Kazakhstan region (2016): In 2016, a case of financial support was registered in the West Kazakhstan region. During the investigation, financial flows aimed at supporting terrorist activities were identified.
2. Money laundering through bank accounts (multiple instances): Various cases of money laundering through bank accounts have been uncovered at different times in Kazakhstan. Money laundering often involves the use of complex schemes, including the creation of fictitious companies and the movement of funds through various bank accounts.
3. Efforts to combat the financing of terrorism and the legitimization of illicit income: Kazakhstan is actively working to strengthen its legal framework and infrastructure to combat

terrorism financing and money laundering. These efforts include implementing measures to freeze the financial assets of terrorist organizations and mandating the reporting of suspicious transactions.

The most vulnerable to money laundering threats are:

– Tax crimes.

– Illegal economic activities.

– Corruption and embezzlement of budgetary funds.

– Fraud.

– Illegal drug trafficking.

Let's make an effort to evaluate potential risks and the likelihood of their occurrence, starting with an optimistic standpoint:

1. Due to various factors, terrorists still favor the use of physical cash. Firstly, terrorists predominantly operate in countries with less advanced technological sectors, making cryptocurrency operations challenging for them. Secondly, the enforcement of customer verification laws and anti-money laundering measures adds further hurdles for terrorists to access cryptocurrencies. Additionally, government agencies have initiated the tracking of transactions on the most popular blockchains. Consequently, opting for physical currency offers a higher level of anonymity and proves to be more challenging to track [3].

2. Additionally, certain terrorist networks have established their own payment systems. All of these factors render the widespread adoption of cryptocurrencies for terrorism financing impractical [4].

3. Another argument from Western experts suggests that terrorists currently lack the essential skills for effectively employing cryptocurrencies. It is believed that utilizing cryptocurrencies demands specialized knowledge in information security. Furthermore, cryptocurrency values are highly volatile, making them less attractive to both regular users and terrorists.

Another challenge for national and international security is the development of shadow marketplaces that maximize anonymity in actions and transactions within the market.

Methodology. Despite the evident need for well-established, science-based anti-money laundering (AML) techniques, methods for detecting money laundering are somewhat limited [5]. The existing on Anti Money Laundering methods falls into two primary categories:

1. Unsupervised Learning: These methods aim to identify data patterns without prior information regarding which data points correspond to money laundering.

2. Supervised Learning: In contrast, supervised learning methods strive to learn patterns that distinguish money laundering from legitimate financial operations. This is achieved by utilizing labeled data where the outcomes (money laundering or not) are known [6].

Supervised learning is generally preferred when there is access to data with known outcomes or labels. However, in the context of anti-money laundering (AML), this poses a challenge. Unlike other forms of financial fraud, financial institutions seldom determine whether a money laundering suspect is definitively guilty of a crime. To address this issue, we can circumvent it by modeling "suspicious" behavior rather than actual money laundering [5]. Machine Learning (ML) belongs to the realm of Artificial Intelligence (AI) applications, and its core purpose revolves around training machines using historical data. This training process equips machines with the ability to comprehend and classify previous datasets, ultimately culminating in the creation of highly effective and accurate prediction algorithms [7]. In this article, we will conduct testing and comparative analysis of the following algorithms. Our paper represents a comparative analysis of machine learning algorithms. The choice of these algorithms was influenced by their successful application in analogous industry challenges.

Here some classification approaches which can be used for money laundering:

a. Logistic Regression: straightforward yet efficient classification algorithm that models the likelihood of a suspicious transaction.

b. Random Forest: A robust ensemble learning technique capable of capturing intricate patterns within transactional data.

c. Support Vector Machines (SVM): Proficient in segregating transactions into categories of suspicion and non-suspicion through the utilization of hyperplanes.

This article will explore the use of a machine learning algorithm, namely K-means clustering and logistic regression in the context of money laundering detection.

Clustering is a data classification technique where data is grouped into multiple categories based on specific features. This grouping ensures that data within the same category exhibit maximum similarity, while data in different categories show minimal similarity [8].

The K-means clustering algorithm is a widely adopted method for clustering data. It involves dividing objects into clusters based on a specified number of clusters. The primary goal is to maximize the similarity among objects within the same cluster while minimizing the similarity between objects in different clusters. This algorithm is known for its simplicity and efficient clustering capabilities. It finds applications in various domains, including data mining, pattern recognition, and image analysis. When applied to stock prediction, it can quickly compute and yield accurate clustering outcomes. However, it has some drawbacks, such as sensitivity to initialization and susceptibility to getting stuck in local extremes.

Here are the steps of the algorithm:

1. Begin with a dataset A containing B objects, where n = 1, 2, ..., m, $A = \{am\}n$ and select i objects randomly as the initial cluster centers.

2. Calculate the distance between the m-th object (*am*) and the j-th cluster center (*cj*) using the formula:

$$D(a_m, c_j) = \sqrt{(a_m - c_j)2} \tag{1}$$

3. Determine the minimum distance *D*min(*am, cj*) from the m-th object (*am*) to the j-th cluster center (ci). Assign objects to the nearest cluster based on the condition:

$$Cj = \{am: D\,(am - cj) < D(am - cz),\ 1 \le z \le i\} \tag{2}$$

4. Compute the mean of objects within the same cluster to update the cluster center:

$$c_j = \frac{1}{n_z}\left[\sum_{\forall A_m \epsilon Y_z} Am\right] \tag{3}$$

where n_z represents the number of objects in the z-th class, and Zj is the subset of all object collections in class j.

5. Repeat steps (2)-(4) until the algorithm converges.

The K-means clustering algorithm typically evaluates the clustering effectiveness using the sum of squared error's function:

$$V = \sum_{z=1}^{i} \sum_{j=1}^{Y_z} \left|a_j^z - c_z\right|^2 \tag{4}$$

where i represents the number of clusters, Y_z denotes the size of cluster z, *aj* represents an object in cluster z, c_z is the cluster center, and $\left|a_j^z - c_z\right|^2$ represents the distance from object *aj* to cluster center c_z .

Nonetheless, the K-means clustering algorithm exhibits certain limitations. To begin with, the selection of the cluster number k relies on human discretion, typically guided by experience. For our research we will choose several clusters 20. Various k values that can yield distinct clustering outcomes. Moreover, the application of different distance calculation techniques may lead to divergent clustering results. Furthermore, the algorithm's objective function is susceptible to convergence into local optima.

Logistic regression models are commonly employed to understand the connection between a qualitative variable, a dichotomous dependent variable (either binary or binomial logistic regression), and one or more independent explanatory variables, which can be of qualitative or quantitative nature. Initially formulated in an exponential form, these models can be transformed into a logarithmic equation (logit), enabling their use as a linear function [9].

Logistic regression is a statistical technique used to predict binary outcomes by considering one or more predictor variables. Its primary goal is to determine whether a variable instance belongs to a specific category. This approach finds applications in various fields, including:

– Assessing credit scores

– Evaluating the effectiveness of marketing campaigns

– Predicting the revenue of a particular product

The predictions generated typically involve outcomes like Yes/No, Alive/Dead, Pass/Fail, and so on.

Logistic regression can accommodate a wide range of features, encompassing both continuous and discrete variables, as well as non-linear features. This technique relies on the utilization of the Sigmoid function, often referred to as the Logistic function.

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{4}$$

Given the nature of logistic regression models, two main types of analyses can be conducted:

1. Assessing the significance of the relationship between each covariate and the dependent variable.

2. Categorizing individuals into the two groups of the dependent variable based on their probability of belonging to either category.

In the context of this study, the second type of analysis is the primary focus. Logistic regression is a valuable statistical tool for estimating individual probabilities.

However, when dealing with a relatively large number of covariates or when these covariates exhibit high correlations, the estimated model parameters may become unstable. Consequently, it becomes necessary to carefully select the variables that will be utilized in training the model.

Data sources. Real financial transaction data is highly limited in access, primarily due to confidentiality and privacy concerns. Even when access is possible, accurately labeling each transaction as either involving money laundering or being legitimate, as mentioned before, poses significant difficulties. Synthetic transaction data provided by IBM presents a way to circumvent these challenges. Real data sets couldn't be used for analysis that's why they have been used from big data set website kaggle.com [10].

The information provided here is derived from a virtual world populated by individuals, businesses, and banks. People interact with one another and with companies, and similarly, businesses engage with both other companies and individuals. These interactions can take various forms, including buying consumer goods and services, placing orders for industrial products, paying salaries, repaying loans, and more. Typically, these financial transactions occur through banks, where both the payer and the recipient have various types of accounts, ranging from checks to credit cards and bitcoins.

A certain (small) segment of individuals and companies in the model engage in unlawful activities, such as smuggling, illegal gambling, extortion, etc. Criminals obtain proceeds from these illicit activities and then try to obscure the source of these illegal funds through a series of financial transactions. These financial maneuvers aimed at concealing illicit funds are referred to as money laundering [10].

The data generator responsible for the information presented here not only imitates illegal activities but also tracks the funds obtained from these unlawful activities across a varying number of transactions. This capability enables the detection of money laundering transactions, even when they are several steps removed from their unlawful origin. Thanks to this underlying framework, the generator finds it more straightforward to categorize individual transactions as either legitimate or illicit.

It's worth emphasizing that this IBM generator models the complete money laundering process:

1. Placement: The introduction of illegitimate funds, such as those from contraband sources.
2. Layering: The commingling of illegal funds within the financial system.
3. Integration: The utilization of these illicit funds.

Moreover, a noteworthy advantage of using synthetic data is that, an actual bank or institution usually has access only to a portion of transactions related to money laundering: those involving that specific institution. Transactions taking place in other banks or between different banks often evade detection. Consequently, models constructed based on actual transactions from a single institution can offer only a restricted perspective of the broader financial landscape [11].

Conversely, these synthetic transactions encompass an entire financial ecosystem. As a result, it becomes feasible to develop money laundering detection models that encompass a broad spectrum of transactions between institutions and apply these models to make assessments specifically concerning transactions within a particular bank [12].

Certainly, money laundering detection is more effectively addressed using classification and anomaly detection methods. Here are the steps and methods typically employed for this purpose.

Results. To assess the outcomes, we employed SAS for algorithm precision and calculations.

The application of K-means clustering for money laundering detection involves the following steps [13]:

1. Feature Engineering: Relevant features were derived from synthetic financial transaction data. These features encompass transaction amount, frequency, source, destination, and timestamp. This dataset was obtained from the Kaggle website.

2. Data Preparation: Data underwent a cleaning and preprocessing phase to ensure it adheres to an appropriate format for analysis. The data was imported into SAS for analysis, leveraging SAS's machine learning and data mining capabilities, which facilitate tasks such as predictive modeling, clustering, classification, and other advanced analytics.

3. Cluster Generation: The K-means algorithm was employed on the preprocessed data. The choice of 20 clusters (K) was based on problem-specific considerations and domain expertise. The algorithm segregated transactions into K clusters by assessing their similarity within the feature space (Fig. 1) [14].

4. Identification of Anomalies: Following the clustering process, transactions were analyzed in smaller, more manageable subsets. Clusters displaying notable deviations from the standard pattern were detected. Transactions within these clusters could be marked as potential anomalies or suspicious occurrences (Fig. 2).

Replace=FULL Radius=0 Maxclusters=20 Maxiter=1

Initial Seeds

Cluster	Timestamp	From Bank	To Bank	Amount Received	Amount Paid	Is Laundering
1	0.128525414	0.000193656	0.696185443	0.000000000	0.000000000	1.000000000
2	0.042815618	0.870121414	0.000039294	0.000000000	0.000000000	0.000000000
3	0.936483620	0.000000000	0.000000000	0.000000011	0.000000013	0.000000000
4	0.008052479	0.999974741	1.000000000	0.000000000	0.000000000	0.000000000
5	0.764710504	0.660978608	0.000000000	0.000000001	0.000000001	0.000000000
6	0.324809490	0.055918855	0.589918971	1.000000000	1.000000000	0.000000000
7	0.048825517	0.713871940	0.713572818	0.000000000	0.000000000	1.000000000
8	0.480791893	0.996244759	0.621064686	0.000000002	0.000000002	0.000000000
9	0.933459031	0.600476562	0.600491730	0.000000008	0.000000009	0.000000000
10	0.007188310	0.699984844	0.005759305	0.000000003	0.000000003	1.000000000
11	0.127347003	0.348008150	0.033393864	0.000000355	0.000000355	0.000000000
12	0.472817975	0.062202850	0.062204422	0.598331346	0.007690239	0.000000000
13	0.480124126	0.000039293	0.719980466	0.000000000	0.000000000	0.000000000
14	0.992143923	0.026297916	0.003151900	0.000000002	0.000000002	1.000000000
15	0.667530835	0.602713428	0.615920043	0.000000004	0.000000004	1.000000000
16	0.112420457	0.000193656	0.000328381	0.000000003	0.000000003	1.000000000
17	0.105389269	0.034745806	0.000005613	0.534581959	0.534581959	0.000000000
18	0.359415508	0.314516899	0.000016840	0.923187540	0.923187540	0.000000000
19	0.817699741	0.000328373	0.699191396	0.000000039	0.000000039	1.000000000
20	0.764710504	0.660978608	0.000000000	0.000000010	0.000000010	1.000000000

Criterion Based on Final Seeds =	0.0761

Figure 1. Initial seeds of 20 clusters on SAS

Cluster Means

Cluster	Timestamp	From Bank	To Bank	Amount Received	Amount Paid	Is Laundering
1	0.226271824	0.030814550	0.490557922	0.000144323	0.000144323	1.000000000
2	0.089951688	0.695890913	0.128310454	0.000003800	0.000003800	0.000000000
3	0.444649135	0.028299958	0.039857838	0.000003987	0.000002559	0.000000000
4	0.035472099	0.659063113	0.663363377	0.000002301	0.000002299	0.000000000
5	0.423409242	0.633911514	0.054216470	0.000001421	0.000001421	0.000000000
6	0.324809490	0.055918855	0.589918971	1.000000000	1.000000000	0.000000000
7	0.174592500	0.504060081	0.490394448	0.000001357	0.000001357	1.000000000
8	0.371725705	0.677885408	0.538654332	0.000002470	0.000002461	0.000000000
9	0.404979823	0.395732495	0.462457637	0.000003645	0.000003605	0.000000000
10	0.205813767	0.503373025	0.049947159	0.000007201	0.000007201	1.000000000
11	0.130751881	0.050624713	0.056948332	0.000005286	0.000003875	0.000000000
12	0.159627622	0.045180774	0.045181915	0.487317452	0.006411136	0.000000000
13	0.276743225	0.040405022	0.517553948	0.000003885	0.000003885	0.000000000
14	0.600128531	0.028205288	0.029975186	0.000001629	0.000001629	1.000000000
15	0.481582469	0.511966074	0.486860855	0.000006539	0.000006539	1.000000000
16	0.263541588	0.027689688	0.033614021	0.000037153	0.000037153	1.000000000
17	0.159627622	0.045180774	0.196061107	0.487317452	0.487317452	0.000000000
18	0.359415508	0.314516899	0.000016840	0.923187540	0.923187540	0.000000000
19	0.545300170	0.034252975	0.490560230	0.000001411	0.000001411	1.000000000
20	0.525680259	0.502375384	0.038844144	0.000002635	0.000002635	1.000000000

Figure 2. Cluster means on SAS

Based on the analysis data and the results of clustering in the figure below, we can identify the nearest cluster and the distance between centroids.

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	591	0.0778	0.3445		19	0.3190
2	173790	0.0902	0.4707		5	0.3472
3	1113313	0.0384	0.4897		11	0.3151
4	182120	0.0938	0.4687		8	0.3591
5	160054	0.0781	0.3922		2	0.3472
6	1	.	0		18	0.6541
7	174	0.0899	0.3418		15	0.3071
8	116294	0.0868	0.5394		9	0.2941
9	111864	0.0778	0.4484		8	0.2941
10	407	0.0726	0.3226		20	0.3201
11	2254690	0.0675	0.4694		3	0.3151
12	5	0.0928	0.3035		11	0.4884
13	961031	0.0918	0.4697		9	0.3818
14	901	0.0479	0.3304		16	0.3366
15	166	0.0914	0.4640		7	0.3071
16	2329	0.0604	0.3100		14	0.3366
17	5	0.1307	0.3743		12	0.504
18	1	.	0		6	0.654
19	290	0.0749	0.3647		1	0.319
20	319	0.0767	0.3915		10	0.320

Figure 3. Cluster summary on SAS

Small datasets encompass a 10-day window of "actual" data, spanning from September 1 to September 10, 2022. It's worth mentioning that the dataset includes only a limited number of transactions occurring after September 10. This occurrence is attributable to the fact that certain money laundering schemes entail multiple days to reach completion. For instance, if an individual initiates a fund laundering process on September 10 that requires an additional 2 days to conclude, the dataset will incorporate transactions related to this money laundering cycle on September 11 and 12.

The second analysis of logistic regression algorithms. A dataset has been taken from kaggle and analyzed by the function “is laundered”. Selection method is a backward method (Fig. 4).

Consider Y as the binary target variable, taking the value 1 when an event occurs and 0 when there's no event. X represents the explanatory input variables, and we refer to the probability associated with them as the "response probability."

Data Set	WORK.IMPORT
Dependent Variable	Is Laundering
Selection Method	Backward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	5078345
Number of Observations Used	5078345

Class Level Information		
Class	Levels	Values
Payment Format	7	ACH Bitcoin Cash Cheque Credit Card Reinvestment Wire

Dimensions	
Number of Effects	4
Number of Parameters	10

Backward Selection Summary					
Step	Effect Removed	Number Effects In	Number Parms In	F Value	Pr > F
0		4	9		

Selection stopped because the next candidate for removal has SLS < 0.05.

Figure 4. Imported data to SAS

The Akaike Information Criterion (AIC) serves as a measure for evaluating the goodness of fit of various regression models (Fig.5).

The Bayesian Information Criterion (also known as Schwarz Criterion or SC) is employed for model selection among a group of parameterized models, each having a different number of parameters.

One key distinction compared to the Akaike criterion is that the Bayesian Information Criterion penalizes the inclusion of additional parameters [15].

In essence, lower values of both AIC and SC indicate a model's superior ability to fit the data. In Figure described fit criteria for is laundering and determined full model.

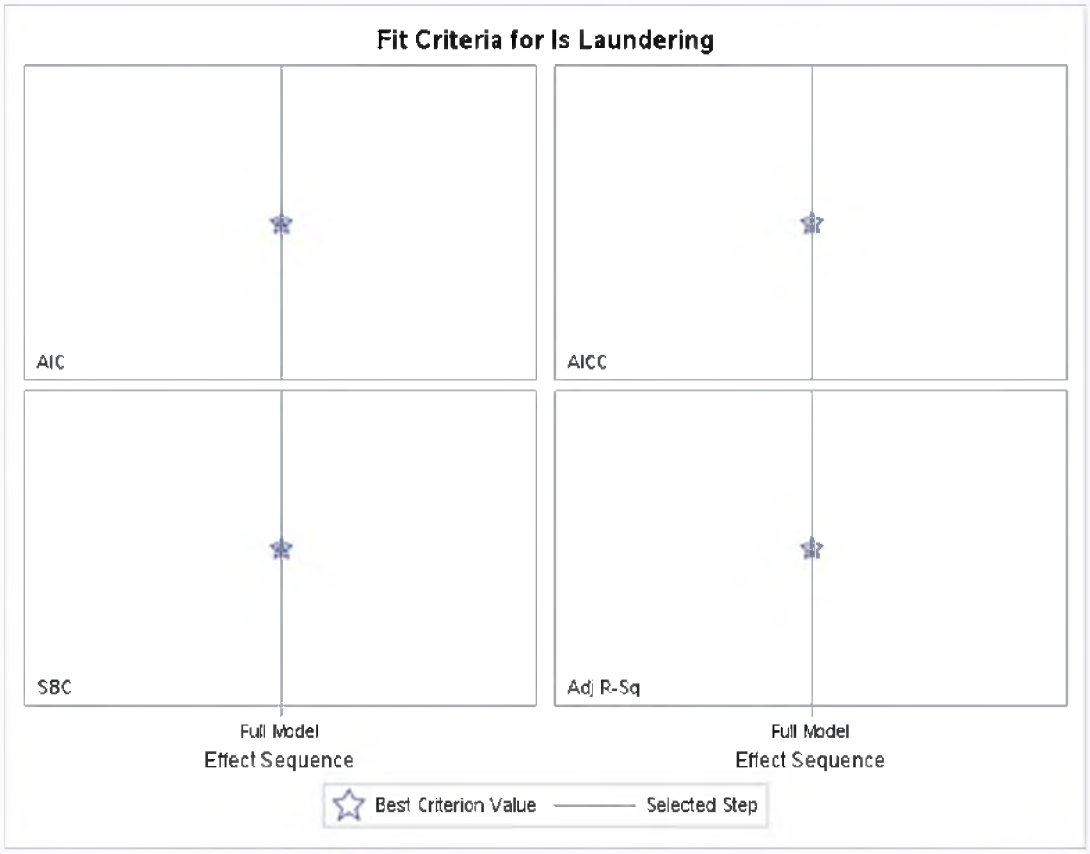

Figure 5. Fit criteria diagram for "Is Laundering" on SAS

Given that the p-value is below 0.05, it indicates that the logistic regression model, as a complete entity, holds statistical significance (Fig. 6).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	28.41269	3.55159	3506.72	<.0001
Error	5.08E6	5143.30974	0.00101		
Corrected Total	5.08E6	5171.72243			

Figure 6. Analysis of "Is Laundering"

Discussion. The discussion revolves around the application of K-means clustering and logistic regression in money laundering detection. The primary focus of this analysis is to assess the efficacy of these machine learning methods in identifying suspicious financial transactions.

K-means clustering is a valuable tool for grouping transactions based on their similarity within the feature space. However, it has certain limitations, including sensitivity to initialization and the risk of converging into local optima. The choice of the number of clusters (K) requires careful consideration and may yield different results for distinct K values. The clustering results reveal anomalies that can be categorized as potential suspicious activities.

In contrast, logistic regression offers a straightforward approach to modeling the likelihood of suspicious transactions. However, in the context of money laundering, where guilt is rarely definitively established, modeling "suspicious" behavior becomes more pragmatic. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) help assess the goodness of fit for different regression models.

Overall, both clustering and regression techniques hold promise in money laundering detection, with the choice depending on specific use cases and data availability. Future research should focus on improving these models and addressing their limitations for more robust anti-money laundering efforts.

Conclusions. The article mentions that Kazakhstan is actively working to strengthen its legal framework and infrastructure to combat terrorism financing and money laundering.

Furthermore, it highlights that the national risk assessment in Kazakhstan is aimed at identifying vulnerabilities in both financial and non-financial sectors, developing measures to minimize money laundering risks, and promoting a unified understanding of these risks among relevant authorities.

In conclusion, this study investigates the application of machine learning techniques, specifically K-means clustering and logistic regression, for money laundering detection. The research aimed to assess the effectiveness of these methods in identifying suspicious financial transactions.

K-means clustering, a data classification technique, showed promise in grouping transactions based on their similarity within the feature space. However, it exhibited certain limitations, including sensitivity to initialization and potential convergence into local optima. The choice of the number of clusters (K) proved crucial and impacted the clustering results.

On the other hand, logistic regression offered a straightforward approach to modeling the likelihood of suspicious transactions, considering binary outcomes. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were employed to evaluate the model fit.

Both clustering and regression techniques provide valuable tools in money laundering detection, with their suitability depending on the specific use case and available data. Future research should focus on refining these models, addressing their limitations, and incorporating real-world financial data to enhance anti-money laundering efforts and minimize false positives and false negatives. Overall, these methods represent critical steps in combating the complex and evolving challenge of money laundering in the financial sector.

Acknowledgements. This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19576825, “Development of machine learning methods and algorithms to identify the financing of terrorist activities in the Republic of Kazakhstan”).

References

1. Alkhalili, M., & Qutqut, M. H. (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.
2. Agency of the Republic of Kazakhstan for Financial Monitoring. (2021). National Risk Assessment of Money Laundering (ML) and Financing of Terrorism (TF).
3. Mohammed, H. N., Malami, N. S., Thomas, S., Aiyelabegan, F. A., Imam, F. A., & Ginsau, H. H. (2022). Machine Learning Approach to Anti-Money Laundering: A Review. In Proceedings of the 4th IEEE Nigeria International Conference on Disruptive Technologies for Sustainable Development, NIGERCON DOI: 10.1109/NIGERCON54645.2022.9803072.
4. Goldman Z.K., Maruyama E., Rosenberg E., Saravalle E., Soloman-Strauss J. (2017) Terrorist Use of Virtual Currencies / Center for a New American Security. May 3. https://www.cnas.org/publications/reports/terrorist-use-of-virtual-currencies.
5. Jullum, M., Løland, A., & Huseby, R.B. (2020). Detecting money laundering transactions with machine learning.Journal of Money Laundering Control. DOI: 10.1108/JMLC-07-2019-0055.
6. Colladon, A.F. and Remondi, E. (2017),Using social network analysis to prevent money laundering,Expert Systems with Applications, Vol. 67, pp. 49-58.
7. M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning. Cambridge, MA, USA: MIT Press, 2012.
8. S. Kapil, M. Chawla and M.D. Ansari, On K-means data clustering algorithm with genetic algorithm, International Conference on Parallel. IEEE, (2017), 202-206.
9. Zhang, Y., Trubey, P. Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. Comput Econ 54, 1043-1063 (2019). https://doi.org/10.1007/s10614-018-9864-z.
10. Smith, J. (2023, October 5). Machine Learning Algorithms for Anti-Money Laundering.Kaggle.https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml/discussion/427517
11. Rakhmetulayeva, S., & Kulbayeva, A. (2022). Building Disease Prediction Model Using Machine Learning Algorithms on Electronic Health Records' Logs. CEUR Workshop Proceedings, 3382, 188-197. ISSN 1613-0073.
12. Becketnova, Y.M. (2020). Analysis of automation possibilities for detecting unscrupulous microfinance organizations based on machine learning methods. Finances: Theory and Practice, 24(6), 38-50. https://doi.org/10.26794/2587-5671-2020-24-6-38-50
13. Domashova, J., & Mikhailina, N. (2021). Usage of machine learning methods for early detection of money laundering schemes. Procedia Computer Science, 190, 184-192. https://doi.org/10.1016/j.procs.2021.06.033.
14. Yang, G., Liu, X., & Li, B. (2023). Anti-money laundering supervision by intelligent algorithm. Computers & Security, 132, 103344. https://doi.org/10.1016/j.cose.2023.103344.
15. Lokanan, M.E. (2023). Predicting money laundering sanctions using machine learning algorithms and artificial neural networks. Applied Economics Letters. https://doi.org/10.1080/13504851.2023.2176435.