# Machine Learning Methods to Detect Terrorist Financing

Aigerim K. Bolshibayeva[1], Sabina B. Rakhmetulayeva[1] and Aliya K. Kulbayeva[1]

[1] *International Information Technology University, Manas St. 34/1, Almaty, 050040, Kazakhstan*

## Abstract

Within the global issue of money laundering, this study conducts an extensive national risk assessment, with a specific focus on Kazakhstan. Advanced methods are employed to identify vulnerabilities in both financial and non-financial sectors and to evaluate the potential risks linked to money laundering. The research utilizes inventive techniques, including both unsupervised and supervised learning methods, to scrutinize patterns in financial transactions, with the goal of differentiating between legitimate transactions and those that may involve money laundering. The use of K-means clustering and logistic regression yields promising outcomes in identifying irregularities and suspicious transactions. Through the incorporation of synthetic financial transaction data, this research provides insights into the concealed nature of money laundering practices. This study represents an initial stride towards improving anti-money laundering endeavors and reinforcing the legal and institutional framework in Kazakhstan. The findings deliver valuable perspectives on the detection of money laundering and its ramifications for both national and international security.

This article presents an overview of the data mining techniques that can be employed to detect financial offenses, including the financing of terrorist organizations. In addition to data preprocessing for a machine learning model to detect financing of terrorist organizations based on publicly available data to ascertain the most significant set of financing anomalies in data.

## Keywords

Machine learning, neural networks, terrorist financing, boosting, classification algorithms

## 1. Introduction

Nowadays, the actions of terrorist organizations, groups, and people are acknowledged as being one of the primary sources of threats to the national security of the Republic of Kazakhstan in the area of security. The extent of the repercussions of terrorist acts and the substantial number of people who are killed or injured as a direct result of their commission are the primary factors that contribute to the high level of this danger.

The degree of financial support, as well as the availability of material and technological resources, has a direct bearing on the severity of terrorist action. In this context, one of the most essential instruments in the battle against international terrorism is the practice of placing a freeze on the assets of terrorist groups and shutting off routes that are used to finance terrorist operations.

The detection systems for suspicious (abnormal) behavior that are utilized by financial institutions nowadays are mostly based on a set of criteria that have been created by specialists in the field. Because these guidelines are not malleable and cannot be readily adapted, they are susceptible to being broken and manipulated in many ways. Another issue is that systems produce a high number of false positives, which are time-consuming to process due to their volume.

Solutions based on machine learning may be continually educated and updated with fresh data, which makes them adaptive and eliminates the need for specialists to develop new rules.

It is possible to continually train and update machine learning-based solutions with fresh data, which makes them adaptive and eliminates the need for consultants to develop new rules. The

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

a.bolshibayeva@iitu.edu.kz (A.K. Bolshibayeva); s.rakhmetulayeva@iitu.edu.kz (S.B. Rakhmetulayeva)
0000-0003-1191-4249 (A.K. Bolshibayeva); 0000-0003-4678-7964 (S.B. Rakhmetulayeva); 0009-0002-7245-8312 (A.K. Kulbayeva)

decision tree (DT) method serves as the foundation for the solution that is suggested in this research. DT is applicable because it is simple to see and comprehend, which enables it to provide a justification for the judgments that it makes. In this study, the decision tree expansions that are being examined are geared at cutting down on the amount of false positives. The drawback of extensions is that they also limit the amount of fraudulent activities that are identified, which results in a trade-off between the number of false positives and the number of fraudulent activities that are not identified.

**The main aim of the research** is the development of a machine learning model to identify the financing of terrorist organizations based on open data to determine the maximum possible set of funding anomalies.

## 2. Literature review

Terrorism, which has a negative impact on the standard of living of people all over the globe, is recognized as one of the most significant dangers facing contemporary civilization. Terrorism's purpose is to sow discord by sowing fear, worry, and insecurity on a broader scale than that of an individual, and its intended effect is to bring about instability. The most recent information from the Global Terrorism Database (GTD) indicates that there were 1,211 separate acts of terrorism carried out throughout the globe in 2019, resulting in 6,362 persons sustaining injuries as a direct consequence of these assaults.[1]

Figure 1 [2] is a graphic representation of the globe map that covers many forms of terrorist activity. According to the information shown by this map, the distribution sites of terrorist activity are situated in the region immediately around Kazakhstan.

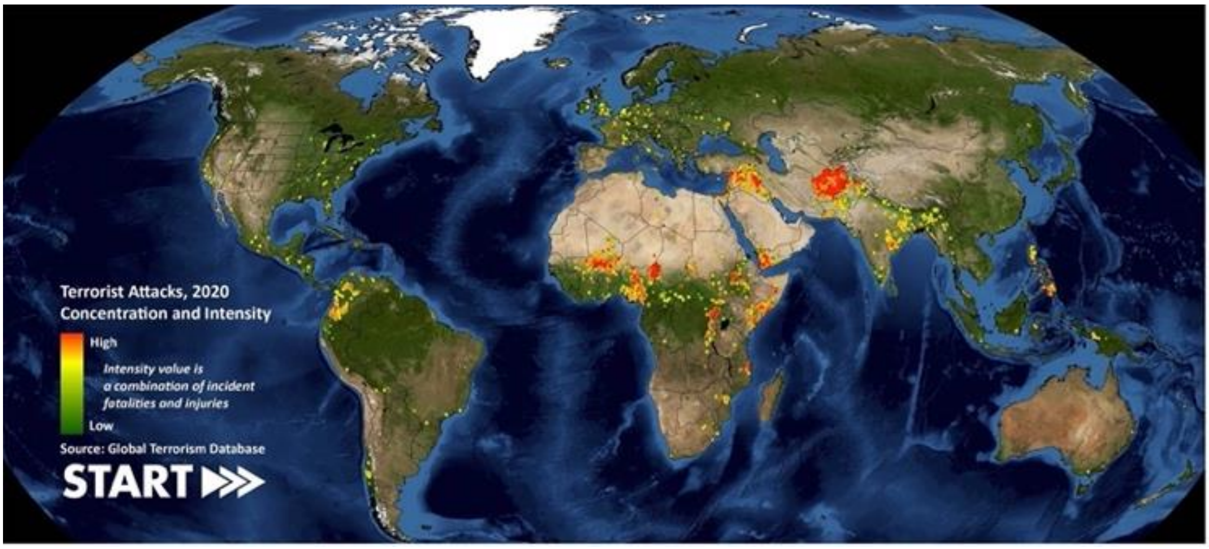

Terrorist Attacks, 2020
Concentration and Intensity
High
Intensity value is a combination of incident fatalities and injuries
Low
Source: Global Terrorism Database
START

**Figure 1**: GTD world map showing the intensity of terrorism in 2020

Because of the impact that terrorism has on the economy, industry, and financial institutions as well as the whole country, it is important and unavoidable that suspicious conduct connected to terrorist funding be identified and investigated. Because of the significant quantity of financial data and the enormous number of transactions that are involved, banks offer a conducive environment for terrorist financing individuals to mask the origin of terrorist funding. As a result, the techniques used for terrorist financing have gotten more complex and harder to track. Because of this, the choice on the detection of terrorist funding has to strike a compromise between precision and processing speed. The most crucial step in solving this problem is to find an appropriate strategy for identifying terrorist funding in financial institutions and banks. This may be accomplished by using an appropriate machine learning methodology with regard to the data set. There is a wide variety of strategies and procedures for identifying irregular money, which makes comparison challenging. "However, it is necessary to conduct an analysis, review, and comparison of various methods for detecting terrorist financing in order to identify crimes, patterns, and unusual behavior as well as money laundering groups for the purpose of financing terrorism" [3].

The application of rule-based approaches followed the introduction of statistical methods and sequence matching in the late 1990s [4]. These methods were first used to identify unusual financial transactions. After some time had passed, financial institutions started integrating statistical and machine learning models into the automated algorithms they used. It is difficult to comply with anti-crime regulations due to the very complex nature of these models, which is caused by the increasing number of client transactions and automated customer interactions. The use of supervised learning techniques, which entail learning from a dataset that has been labeled and then modeling the categorization of incoming data into several label categories, has been suggested as a potential solution to the issue of huge dataset sizes in some of the most current research that has been conducted. Therefore, supervised learning algorithms are only able to identify potentially suspicious behaviors, patterns, and transactions if they are comparable to the data that they were trained on.

While a complicated transaction mechanism happens, experts must retry the diagnostic method, with the potential consultation of a narrow specialist. This slows down the prompt discovery of abnormal financial processes and may need the consultation of a narrow specialist. Additionally, the transaction research sets that are used by experts eventually go out of date and are often unable of adapting to new circumstances.

Utilizing ML approaches is one way to get over the obstacles discussed earlier in this paragraph. The use of neural networks has become more popular as a solution to a variety of data categorization, logging, and error-checking issues. In our proposal to design a system for the detection of financial criminal actions of terrorist funding, we propose to make use of neural networks as the primary technique of data assessment and annotation, in conjunction with traditional machine learning (ML) approaches.

Due to the significant role that it plays in both the fight against cybercrime and the maintenance of a healthy economy, there is a wealth of research available on the subject of identifying instances of financial fraud.

Researchers commonly utilize outlier detection techniques [5] with severely skewed datasets while attempting to uncover instances of financial crime. Fraud in its many forms may also be committed against financial institutions. According to one study [6,] there are four different types of financial fraud: fraudulent financial reporting, fraudulent transactions, fraudulent insurance, and fraudulent credit.

The research paper titled "A machine learning approach for terrorist financing detection" was written by Raghavan, V., and Rakesh, V., and it was published in the year 2019 in the International Journal of Computational Intelligence and Informatics. This work investigates the use of machine learning techniques to identify the sources of funding for terrorist operations.

The purpose of this study is to build an effective model for the use of machine learning in the identification of financial dealings that are associated with terrorist operations. The authors recommend employing machine learning approaches, such as classification and clustering algorithms, to examine financial data in order to detect typical patterns and anomalies that are connected with terrorist funding.

The article provides a description of a technique that evaluates and contrasts a number of algorithms, as well as preparation of data, selecting and tweaking of machine learning models, and assessment of different algorithms. The authors identify prospective instances of terrorist financing by using a variety of machine learning methods, such as decision trees, random forests, and support vector machines, to categorize monetary transactions as "normal" or "suspicious" in order to locate possible funding sources for terrorist organizations.

Because on the findings of the research, the authors suggest a model for machine learning that has a high success rate in recognizing financial transactions that are associated with terrorist activities. In addition to this, comparisons are drawn with other methodologies already in use, and the authors show why their methodology is preferable.

The paper "Detecting terrorist financing using deep learning-based anomaly detection" was written by Yoon, B., Ahn, J. H., Kim, J., and Lee, D. H., and it was presented at the 18th International Conference on Machine Learning and Applications (ICMLA) in 2019, which was organized by IEEE. The paper was also published in the conference proceedings.

Deep learning and anomaly detection are going to be combined in this project so that a system may be developed to identify financial support for terrorist activities. Deep neural networks are a kind of artificial neural network that are proposed by the authors as a method for analyzing financial data and locating abnormal patterns that are related with terrorist funding.

In the study, a methodology is outlined that involves data preparation, the development of a deep learning model, and the training of the model via the application of labeled data. In order to identify abnormalities in financial data, the authors make use of a variety of different designs for deep neural networks, such as convolutional neural networks and recurrent neural networks.

The authors perform tests and assess the usefulness of the suggested strategy by using data sets linked to the funding of terrorist operations. According to the findings of the research, the use of deep learning in conjunction with anomaly detection may be a useful method for the identification of financial transactions that are affiliated with terrorist groups.

There have been many different approaches investigated for use in the detection of financial fraud. In the field of automobile insurance, Phua et al. [7] employed neural networks, naive bayes, and decision trees to identify instances of fraudulent activity. Another paper combined SVM, genetic programming, logistic regression, and neural networks in order to identify financial reporting fraud of Chinese enterprises. Ravisankar et al., (2011) was the first to do so. The detection of fraudulent activity made use of density-based clustering [8] and cost-sensitive decision trees. Sorournejad et al., [9] cover supervised and unsupervised machine learning based methodologies, such as clustering, artificial neural networks (ANNs), support vector machines (SVMs), and hidden Markov models (HMMs). Wedge et al., [10] handle the issue of unbalanced data, which results to a relatively large number of false positives. Additionally, other studies provide strategies to overcome this problem.

In spite of this, there is a paucity of research on the subject of identifying the financial dealings of terrorist groups, perhaps as a result of the recentness of technological advancements.

The article by Albashrawi and colleagues [11] provides a comprehensive analysis of the techniques that are the most often used to the detection of financial fraud (Table1).

**Table 1**
**Analysis of the techniques that are the most often used to the detection of financial fraud**

| Technique | Frequency of use |
|---|---|
| Logistic Regression | 13% (17 articles) |
| Neural Networks | 11% (15 articles) |
| Decision Trees | 11% (15 articles) |
| Support Vector Machines | 9% (12 articles) |
| Naïve Bayes | 6% (8 articles) |

The use of an accurate representation of the data is one of the distinguishing characteristics of the model that has been presented. Models that are able to enhance this representation have been and continue to be the primary avenues of advancement in deep learning.

The following methods, which were found based on the findings of the review, have been recognized as ways to enhance machine learning algorithms for the detection of anomalous financing:

- The efficiency of machine learning algorithms may be considerably improved by collecting data that is both more diversified and up to date than it currently is. This data should focus on financial transactions and connections to terrorist operations. Obtaining data from a variety of sources, such as banking and financial institutions, law enforcement agencies, international organizations, and public information sources, may be part of this process.

- Applying Convolutional Neural Networks and Deep Learning: Processing complicated financial data and locating previously concealed patterns and anomalies may be made easier with the use of powerful methods like deep learning and neural networks. The capability of models to identify terrorist funding may be improved by using more complex neural network structures and methodologies, such as convolutional neural networks and recurrent neural networks.
- The incorporation of contextual information, such as social affiliations, location, and event data, may assist in better comprehending and predicting the connections that exist between terrorist operations and financial transactions. When attempting to identify shifting patterns of terrorist funding, it is essential to take into account the dynamics of the situation over time.
- Functions of losses and metrics for estimating them: Developing machine learning algorithms involves a number of steps, one of the most crucial of which is selecting the suitable loss functions and estimating metrics. It is essential to take into consideration the particulars of the job of detecting terrorist funding and to make an effort to reduce the number of false positives (also known as false positives) and false negatives (also known as omissions).

## 3. Aim and research question

The main hypotheses of the study:

- Can we define a reliable methodology for the analysis and detection for various scenarios in the absence of labels or reliable data?
- Is it possible to extend the methodology in the presence of labels and generalize well even in the presence of unbalanced classes?
- How can we evaluate the quality of synthetic data?
- Can we improve unbalanced classification with synthetic data?
- There are three main directions for solving the hypotheses of the research:
- Study the literature on the identification of terrorist financing and understand various aspects of the problem.
- Solve the problem of detecting financial fraud on a publicly available set of data samples using controlled machine learning methods.

Following these stages will result in the construction of a framework that incorporates the ideas of analytics and machine learning in order to address fundamental issues.

In this investigation, we begin by defining supervised machine learning algorithms. These are programs that are able to recognize patterns in the data that link features (a quantifiable aspect of the data) with labels (a specific aspect of the data). Learning takes place when algorithms search for patterns by analyzing samples for which labels are already known.

A model, which is an approximation of the underlying relationships between features and labels, is the product of this process. During the test phase, the model assigns predictions to samples that were not part of the training phase, and then we compare these assignments to the labels that were already known.

The purpose of this test is to see how effectively the model generalizes to scenarios that have not yet been encountered. A supervised machine learning issue is referred to as a classification when each label represents one choice from a set of potential classes. On the other hand, a supervised machine learning problem is referred to as a regression when the labels are continuous values.

An earlier technique known as Gradient Boost Machines [13.] has been modernized and given the name XGBoost [12], which is a well-known supervised learning tool in the data science field. XGBoost is a rapid version of the Gradient Boost Machines algorithm.

An ensemble is a model that takes the findings of many other models and mixes them in order to compensate for the deficiencies of each individual model. The majority of the options to replace this process may be categorized as either bagging or boosting [14].

The boosting method is an ensemble of consecutive models, each of which corrects the mistakes made by the model that came before it in the series. In particular, the term "gradient boosting" refers to a method for reducing the amount of prediction residuals that is based on gradient descent. The decision trees that are used are the default and most popular ensemble elements for XGBoost. These elements have a significant variance that may be corrected by boosting. XGBoost, like many other machine learning models, generates outputs that indicate the likelihood that a given sample belongs to a certain class. The choice for a binary classification comes from splitting the expected probability into two halves using a threshold of 0.5 as the dividing line.

Following that, we examined the mechanism for detecting anomalies. Using this technique, you may identify instances within a sample that exhibit abnormal behavior and isolate them for further investigation. This is the primary objective of statistical analysis and machine learning, and there is a wide variety of effective approaches. The primary response is to perform statistical tests in order to model the distribution function of the data in a frequency interpretation or to use Bayesian interpretation in order to analyze the likelihood of sampling the input data while it is inside the modeled probability distribution function. Both of these approaches are viable options. The straightforward method for detecting anomalies is to take each variable and compute the z-score for each sample. This score is connected to the point's distance from the mean, which is quantified in standard deviations.

The difficulty is that if you define an anomalous score based on each parameter independently, you are ignoring the intricate relationships that occur between features. There is a multi-dimensional version of this approach that makes use of the Mahalanobis distance; however, this extension makes the assumption that the distribution of the data is normal, which is not always the case. The Isolation Forest (IF) technique [15] is predicated on the hypothesis that a person or entity should be considered anomalous if it is simpler to distinguish them from the rest of the group in a random subdivision of the feature space. A random measurement is taken after a sample is selected from a larger data collection to get things started. The sample is then divided in half using a random value that falls somewhere within the allowed range of that dimension. The first node of the tree is generated by the algorithm once the user specifies the size and the split point. Additional nodes are generated for the subsamples in a cyclical manner until it is either impossible to do a split or an arbitrary depth of the tree has been achieved. In this tree, a point that is closer to the root node correlates to a situation that is more likely to be isolated; yet, it is possible that this occurrence is the result of random chance. After that, the algorithm will proceed to repeat the whole process of creating trees using fresh samples until the required number of trees has been generated. As a last step, it determines the anomaly score by computing the typical length of the traversal route over all of the trees.

An analysis of variance strategy will be used in order to determine the parameters and attributes of the neural network before moving on to the next step of the process, which is the selection of the machine learning models.

## 4. Carrying out the experiment

The initial step in the preparation phase involves screening faulty values from the dataset and removing rows from the dataset that include those rows. Experts in the field need to provide the format that is anticipated for each variable. Errors in the software, the programming, or the input of the data may all lead to invalid samples. In certain instances, criminals would intentionally forge erroneous information in order to get around different safeguards that are in place. As a result, invalid transactions need to be labeled as suspicious at an early level of the pre-processing stage and filtered out of the remainder of the investigation.

After that, the user has aggregated a number of transactions, and a collection of vectors is generated as a consequence of this process. The majority of cons may be identified by suspiciously high transaction and referral rates, consistent and constant amounts of activity, and extremely short pauses between activities. Recency, frequency, and monetary variables, sometimes known as RFM variables, have the ability to capture this behavior and are commonly employed in various fraud research situations [16]. Any user who has a value in one of the RFM variables that is out of the ordinary should be viewed as potentially malicious. However, uncommon combinations of the values of the variable are more difficult to identify. When RFM variables are computed from transactional data, the resulting data for each account is a collection of time series. The number of transactions that are associated with an account determines how long the time series that represents the account will be. The aggregation of RFM variables for each user is required in order to compare vectors of constant size, but doing so results in an unavoidable loss of information. We suggest providing each time series with many statistical functions all at once, such as the median, mean, or standard deviation. This is something that we propose doing.

We train the anomaly detection algorithm by running it on the dataset that we have been given and using the machine learning program's implementations. In each of these cases, we will need to establish some model parameters before we can proceed. We put the algorithm through its paces with a number of different parameter values and then display those results against the Bayesian Information Criterion (BIC) [17] for each model. This measure cannot be interpreted on its own, however, when comparing the BIC of the two models, it is preferable to have lower values. By locating the point at which the slope of the curve stops noticeably decreasing, we are able to identify the number of components involved.

Every user receives a score based on the algorithm that they utilize. These numbers are measured on a variety of scales, making it impossible to compare them without further processing. First, we scale the output of each algorithm such that the individual in the dataset who seems the most suspicious receives a score of one and the person who seems the least suspicious receives a score of zero.

Logistic regression (LOG), linear and quadratic discriminant analysis (LDA, QDA), least square support vector machines (LS-SVM), decision trees (C4.5), neural networks (NN), nearest-neighbor classifiers (k-NN10, k-NN100), a gradient boosting algorithm, and random forests are the statistical methods that will be utilized in this paper. We are particularly interested in the power and applicability of the random forest and gradient boosting classifiers, both of which have not been substantially examined in the context of credit scoring.

The area under the receiver operating characteristic curve (also known as AUC) will be used to assess each approach. According to Baesens and colleagues (2003), this is a measure of the discriminating capability of a classifier that does not take into account class distribution or the cost of misclassification.

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 dataset A containing B objects, where n = 1, 2, ..., m, $A = \{a_m\}n$ and select i objects randomly as the initial cluster centers.
2. Calculate the distance between the m-th object ($a_m$) and the j-th cluster center ($c_j$) using the formula:

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

3. Determine the minimum distance $D_{\min}(a_m, c_j)$ from the m-th object ($a_m$) to the j-th cluster center (ci). Assign objects to the nearest cluster based on the condition:

$$C_j = \{a_m\text{:}\ D\ (a_m - c_j) < D(a_m - c_z),\ 1 \leq z \leq i\} \tag{2}$$

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, $a_j$ 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 $a_j$ to cluster center $c_z$ .

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.

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.

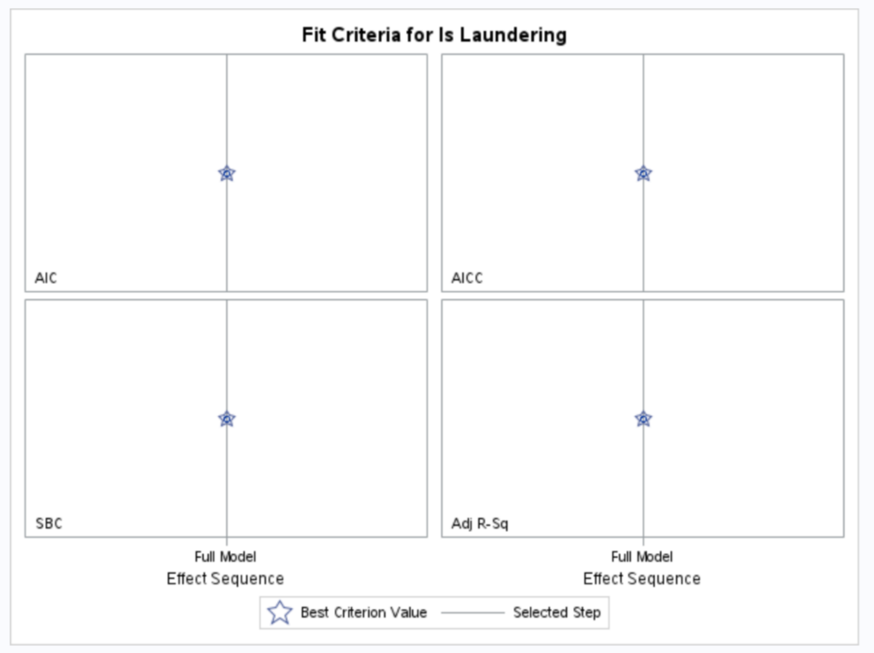
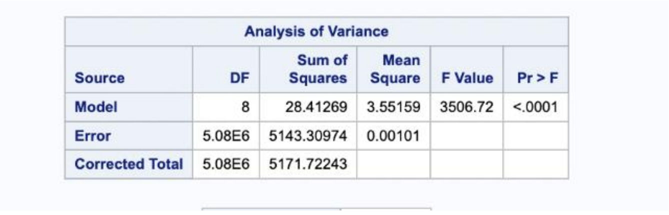


**Figure 2:** Fit criteria diagram for "Is Laundering" on SAS

Given that the p-value is below 0.05, it indicates that the logistic regression model, as a complete entity, holds statistical significance (Fig.3).



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 3:** Analysis of "Is Laundering"

## 5. Conclusion

This article is the first in a series of articles on this study.

This article summarizes a review of mining methods that can be used to detect financial crime and also identifies ways to improve machine learning approaches.

The article discusses Kazakhstan's active efforts to bolster its legal framework and infrastructure in the fight against terrorism financing and money laundering. It also emphasizes that the national risk assessment in Kazakhstan aims to pinpoint vulnerabilities in financial and non-financial sectors, create measures to reduce money laundering risks and promote a common understanding of these risks among relevant authorities.

In summary, the research explores the application of machine learning techniques, specifically K-means clustering and logistic regression, for detecting money laundering. The study's goal was to evaluate the effectiveness of these methods in identifying suspicious financial transactions. K-means clustering, a method for categorizing data, showed promise in grouping transactions based on their similarities within the feature space. However, it had certain limitations, such as sensitivity to initialization and the potential to get stuck in local optima. The choice of the number of clusters (K) was found to be critical and influenced the clustering results.

On the other hand, logistic regression provided a straightforward approach for modeling the likelihood of suspicious transactions, particularly with binary outcomes. The study used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to assess the model's fit.

Both clustering and regression techniques offer valuable tools for detecting money laundering, and their suitability depends on the specific use case and available data. Future research should focus on refining these models, addressing their limitations, and integrating real-world financial data to improve anti-money laundering efforts and reduce false positives and false negatives. In conclusion, these methods are essential steps in the ongoing battle against the complex and ever-changing problem of money laundering in the financial sector.

## 6. 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).

## 7. References

[1] R. Alhamdani, M. Abdullah, and I. Sattar. (2018). Recommender system for global terrorist database based on deep learning. International Journal of Machine Learning and Computing, vol. 8, pp. 571–576.

[2] Source https://www.start.umd.edu/gtd/.

[3] Labib, N. M. and Rizka, M. A. (2020) "Survey of Machine Learning Approaches of Anti-money Laundering Techniques to Counter Terrorism Finance". Springer Singapore. doi: 10.1007/978-981-15-3075-3.

[4] Phua, C. et al. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. doi: 10.1016/j.chb.2012.01.002.

[5] Jayakumar et.al. (2013). A New Procedure of Clustering based on Multivariate Outlier Detection. Journal of Data Science. 11: 69-8.

[6] Jans et.al. (2011). A Business Process Mining Application for Internal Transaction Fraud Mitigation, Expert Systems with Applications. 38: 13351–13359.

[7] Phua et.al. (2004). Minority Report in Fraud Detection: Classification of Skewed Data. ACM SIGKDD Explorations Newsletter. 6: 50-59.

[8] Dharwa et.al. (2011). A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction, International Journal of Computer Application. 16: 18-25.

[9] Sorournejad et.al. (2016). A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.

[10] Wedge et.al. (2018). Solving the False Positives Problem in Fraud Prediction Using Automated Feature Engineering, Machine Learning and Knowledge Discovery in Databases, pp 372-388.

[11] Albashrawi et.al. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015, Journal of Data Science. 14: 553-570.

[12] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree-boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794.

[13] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232.

[14] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[15] Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008a). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, pages 413–422. IEEE.

[16] Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons.

[17] Aggarwal, C. C. (2016). Outlier analysis. Springer, 2nd edition.

[18] IBM; AMLSim; GitHub [online]; updated on: 23.10.2022; Available from: https://github.com/IBM/AMLSim#amlsim.

[19] LOPEZ-ROJAS, Edgar, ELMIR, Ahmad, AXELSSON, Stefan; PaySim: A financial mobile money simulator for fraud detection; In 28th European Modeling and Simulation Symposium, EMSS, Larnaca; Dime University of Genoa, 2016; p. 249-255.

[20] CHEN, Zhiyuan et al. (2018,). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review; Knowledge and Information Systems [online]; 57.2: 245- 285 [accessed on 18.09.2022]; ISSN 0219-3116; Available from: https: //link.springer.com/article/10.1007/s10115-017-1144-z.

[21] Rakhmetulayeva, S., Syrymbet, Z. (2022). Implementation of convolutional neural network for predicting glaucoma from fundus images, Eastern-European Journal of Enterprise Technologiesthis link is disabled, 6(2-120), pages 70–77 DOI:10.15587/1729-4061.2022.269229.

[22] Mukasheva, A., Rakhmetulayeva, S., Astaubayeva, G., Gnatyuk, S. (2022). Developing a system for diagnosing diabetes mellitus using bigdata, Eastern-European Journal of Enterprise Technologiesthis link is disabled, 5(2-119), pages 75–85, DOI:10.15587/1729-4061.2022.266185.

[23] Rakhmetulayeva, S., Kulbayeva A. (2022). Building Disease Prediction Model Using Machine Learning Algorithms on Electronic Health Records' Logs, CEUR Workshop Proceedingsthis link is disabled, 3382.