Transactions on Machine Learning and Artificial Intelligence - Vol. 10, No. 4
Publication Date: August, 25, 2022
DOI:10.14738/tmlai.104.12854.
Deussom Djomadji, E. M., Matemtsap Mbou, B., Tchagna Kouanou, A., Ekonde Sone, M., & Bayonbog, P. (2022). Machine Learning-Based Approach for Designing and Implementing a Collaborative Fraud Detection Model Through CDR and Traffic Analysis. Transactions on Machine Learning and Artificial Intelligence, 10(4). 46-58.

Machine Learning-Based Approach for Designing and Implementing a Collaborative Fraud Detection Model Through CDR and Traffic Analysis

DEUSSOM DJOMADJI Eric Michel
Department of electrical and electronic engineering
College of Technology, University of Buea
Division of information and communications technology
NASPT, University of Yaoundé I

MATEMTSAP MBOU Bequerelle
Division of information and communications technology
NASPT, University of Yaoundé I

TCHAGNA KOUANOU Aurelle
Department of computer engineering, College of Technology
University of Buea, Cameroon
Division of information and communications technology
NASPT, University of Yaoundé I

Michael EKONDE SONE
Department of electrical and electronic engineering
College of Technology, University of Buea

BAYONBOG Parfait
Division of information and communications technology
NASPT, University of Yaoundé I

ABSTRACT

Fraud in telecommunications networks is a constantly growing phenomenon that causes enormous financial losses for both the individual user and the telecommunications operators. We can denote many researchers who have proposed various approaches to provide a solution to this problem, but still need to be improved to ensure the efficiency. Detecting fraud is difficult and, it's no surprise that many frauds schemes have serious limitations. Different types of fraud may require different systems, each with different procedures, parameter adjustments, database interfaces, and case management tools and capabilities. This article uses the K-Means algorithm to handle fraud detection based on Call Detail Record (CDR) and traffic analysis in a telecommunication industry. Our algorithm consists to compare traffic and CDR generated in the network and check if there is abnormal behavior and if yes, our model is used to confirm if users suspecting of fraud are really fraudster or not. To build our model we used real word CDR data collected in November 2021. Our model associates the Differential

Privacy model to encrypt users' personal information, and the k-means algorithm to group users into different clusters. Those clusters represent non fraud users having similar characteristics based on criteria used to build the model. Users having abnormal behavior that can be assimilated to fraudsters are those who are far from the different clusters center. Thanks to a representation in a plan, we better visualize user's behavior. We validated our model by evaluating our segmentation method. The interpretation of the results shows sufficiently that our approach allows to obtain better results. Our approach can be used by all telecommunications operator to reduce the impact of fraud on internet services.

Key Word: Fraud, Traffic, Personal Data, CDR, Machine Learning, K-Means

INTRODUCTION

Mobile Networks are getting smarter day by day and, mobile phone devices such as tablets, smartphones, wearable devices are increasing rapidly [1,2]. Fraud is a world problem. Frauds affect the reputation and the performance of telecommunications companies and leads a huge loss of revenue that affects the reputation and performance of telecommunications companies. Telecommunication fraud also affects the robbery of services and deliberate misuse of data and voice networks. According to [3], Fraud is an intentional fraudulent or dishonest act by one or more individuals, usually for financial gain. In telecommunication companies, time stamps, Call Detail Record (CDR) and traffic data are the element to be considered when searching for traffic anomalies [4]. There are various techniques available for managing and detecting tecommunication fraud these include the use of a fraud management system based on different models, the implementation of network control measures such as the verification of the limits of the invoicing system and the control of customer credit security, the use of DPI (Deep Packet Inspection), and security firewalls [5]. However, these security measures have many limitations in terms of real-time fraud detection and the deployment of corrective measures. Today, Anomaly and fraud detection is an important data analysis and ML task which is helpful for recognizing the network intrusions [6]. ML is the combination of various computer algorithms and statistical modeling that enables computers to perform tasks without being hard-coded [7, 8]. A lot of research in the domain of fraud and anomaly detection in telecommunication industries are use ML. In this paper, we proposed a new architecture based on ML to detect fraud in by using CDR and traffic data. In ML, we used a K-Means algorithm as our unsupervised learning algorithm to build a collaborative fraud detection for a telecommunication company.

The rest of this paper is structured as follow: the state f the art is presented in section 2. We design our architecture and present our technic and method in methodology section. The results and discussions of outcome of this work is presented in section 4. This work ended with a conclusion and future orientations.

STATE OF THE ART

Fraud in the telecommunications sector is increasing every year causing significant financial losses for telecommunications operators and the individual user. Several solutions are proposed in order to reduce or even eliminate the occurrences of fraud in telecommunications networks. However, fraudsters are constantly developing methods to counter the various security measures deployed. Most solutions are based on the analysis of

URL: http://dx.doi.org/10.14738/tmlai.104.12854

CDRs because they contain all the information relating to a user and his behavior on the network. Therefore, several problems arise, namely:
The preservation of user privacy: we plan to develop a solution based on CDRs which contain a lot of private and non-private information. This raises the issue of the collection, processing and security of personal data. In many states, governments at the regional or national level have taken measures to protect personal data.

CDR data processing: CDR containing records of customer conversations such as destination and source number, duration of calls, call start time, and, the telecommunication operator can be a source of information to use in fraud detection [9]. CDR are metadata about how a phone number uses the cellular network [10]. CDR records represent a large amount of data which makes it difficult to handle. Since we need to use a considerable data sample to design our model in order to guarantee its efficiency, then it is essential for us to identify an appropriate method to process this large amount of data.

The scalability of the proposed solution: The telecommunications sector is a sector with an exponential scalability; every day, some new services and software are proposed. It is therefore essential today to carry out a technological watch, autonomous or not, to guarantee the effectiveness of the solutions we proposed to avoid security breaches and new techniques which can lead to the exploitation of the network for unhealthy purposes. As part of this work, we want to propose a solution that can protect the network from any attempt of fraud, known or not.

Nowadays, there might emerge two victims in Telecommunication Fraud [11]:

- Carrier - Since fraudster use carrier resources, it is Since these resources can be used on legitimate costumers, they are immediately out of function.
- Customer - If he falls victim to a fraudulent attack, the integrity of the customer will in turn be challenged to determine if it really is a fraud scenario. If the costumer is really the victim, it's wrong the alert would jeopardize the customer's past trust in the carrier. There's some left Customers can even pay for services used by fraudster, thus being destroyed to further deepen the customer's trust in the carrier.

One of the hottest topics of the decade is machine learning (ML), a subset of artificial intelligence. More and more companies are looking to invest in ML to improve their activities and services. Many investigations have been done in this field to develop methods and technics to solve fraud problem in telecommunication companies. Rehahi et al in 2011 discussed about the fraud problem in VoIP telecommunication networks and evaluated the related existing solutions [5]. Hoffstadt et al in 2014 developed a Framework for fraud detection in a network [12]. Their framework called SUNSHINE offered an association of firewalling and intrusion detection, a CDR analysis, a distributed sensing system, based on AI and statistical analysis [12]. Zhao et al in 2018 investigated to fraud detection in telecommunication by using contents of a call [13]. Indeed, they focused on the implementation of an Android app by using voice recognition technology along with Natural Language Processing to identify the content of calls and justify fraudulent calls based on its capabilities before warning users. They technique weren't efficient because all the customers should have an Android phone. Also in 2018, Subudhi and Panigrahi presented a novel hybrid fraud detection method in mobile telecom networks that identified the normal and malicious

calls [14]. They used genetic algorithm and fuzzy C-Means to construct their model and relied on historical call records [14]. Besides their model, their obtained accuracy remains less than 95%. Sultan et all investigated also in telecommunication fraud detection by using K-Means algorithm [15]. They based on CDR to detect anomalies and used K-means to verify the anomalies and proceeded to resource distribution as well as fault avoidance and detection. They evaluate the MSE of their model and use ARIMA method to predict the future traffic for user. Gbadoubissa et al in 2020 Investigated on an optimized scheme for costs of maintenance operations on cell sites and based on k-means clustering algorithm for allocating field technician to a pool of cell sites [16]. Furthermore, their proposed an initialization method based on sphere geometry to reduce the sensitivity of k-means during initialization.

Deep learning can also be used to detect fraud in CDR and mobile traffic, it is the case of Chouiekha and El Hajb based on deep learning along with Convolutional Neural Network and represented each costumer by an image containing his behaviour (calls, sms, etc.), to uncover more insights into the fraud detection in mobile communications [9]. Trinh et al., implemented a DL algorithm in a semi-supervised manner to detect and predict potential traffic anomalies, such as those caused by unexpected crowds [17]. However, sometimes, we need to convert costumer data in images data before training, this takes time and computing resources that are not efficient for a telecommunication company.

In order to ensure revenue security among telecommunications operators, we therefore ask ourselves the question of How to set up an efficient and high-precision collaborative fraud detection model that takes into account the respect of the privacy of user's personal information in a constantly changing environment?
It is therefore a question for us of designing a model that can:

- ✓ Distinguish unambiguously the profile of a user (fraudster or not) without exposing his personal information;sss
- ✓ Detect an occurrence of fraud with a reliability of at least 95%;
- ✓ End the session of the fraudulent user.

METHODOLOGY

To design our collaborative model that detects fraud on data services, we will use Machine Learning through the K-Means clustering machine learning algorithm to analyze CDRs. Since CDR data is not labeled (fraudster data and non-fraudster data), clustering is the best approach to group users with similar behaviors. This algorithm has many advantages; it is simple, fast, easy to understand and above all, gives a result with high precision. Additionally, Differential Privacy (DP) will be applied to ensure user privacy as we will use data from a real sample to build and evaluate our model. Our model architecture is presented in Figure 1.

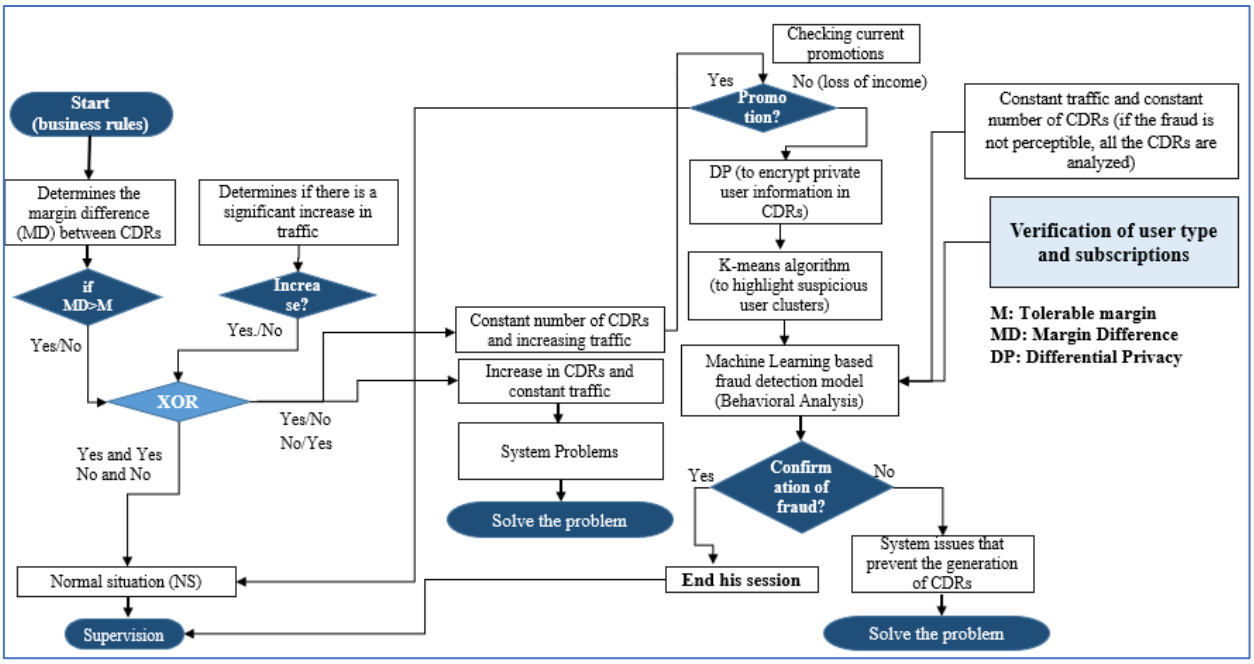


Figure 1: Our Proposed algorithm

Based on the operation of billing systems and data service billing, this algorithm analyzes whether there is an abnormal increase in traffic or CDRs; if so, it shows the CDRs of the users causing the difference and analyzes them thanks to the model to know if they are fraudsters or if it is a system problem.

We have summarized our method in three phases: a data collection and preparation phase, a model construction and training phase and a results analysis phase.

A. Data collection and preparation

The aim of this phase is to collect our data, understand, analyze it and highlight information that will be used for the construction of our model. Regarding the collection, we have collected data from the billing system of a known operator. We obtained two real-world datasets in .csv format for the PCEF and the GGSN dated November 19, 2021. The PCEF one contains 50 rows and 115 columns while the GGSN contains 50 rows and 38 columns. For the preparation, firstly we visualize all the columns inside those files; we have as example, DATE_KEY, RECORDTYPE, SERVEDIMSI, CHARGINGCHARACTERISTICS, USERLOCATIONINFORMATION and many others. We used the ***dropna*** python function to clean our data chosen to use best columns for the construction of our model because they allow us to better describe a user's session for data services. In Figure 2, some our columns can be visualizing

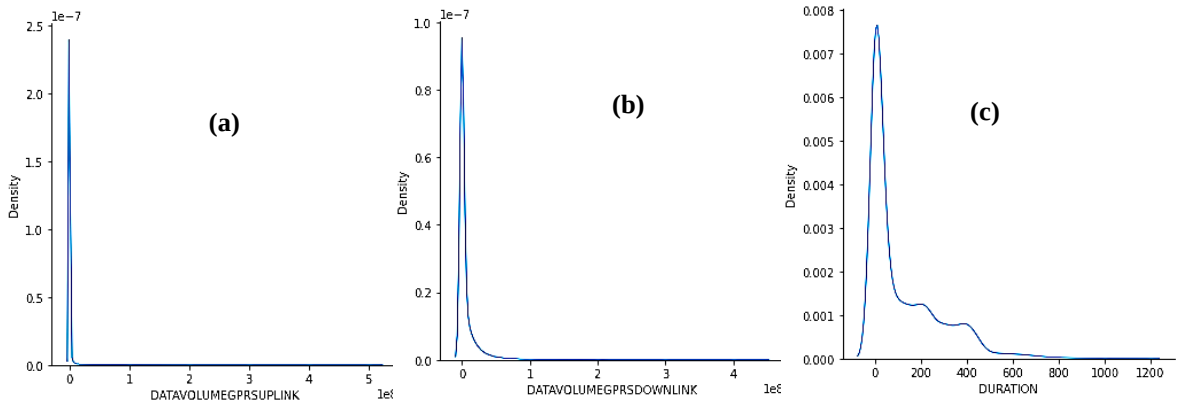


Figure 2: (a) volume of data consumed in uplink function (b) Data volume Downlink function (c): Duration

In figure 2, we can notice that the data distribution function in down Link reveals that the average of the data for the day is around 0.1 * 10^8 bytes, and per user, we have an average duration around 50 (duration unit).

B. Model building

In thus subsection, we describe sub-models, algorithms, and methods used to constitute our own model.

1. The Differential Privacy (DP) Model

As our CDR data is stored in databases thus the security of a statistical database must be ensured. Dalenius [18] described a desideratum that no personal information should be learned without accessing the database. However, Dword [19] later proved that this is impossible; but he proposed a new model called Differential Confidentiality (DP) to ensure that any disclosure whatsoever will be, within a small multiplicative factor, just as likely as the individual participant in the database; in other words, the presence of the individual data would not be the cause of the disclosure of information. The rigorous definition of DP is shown in the following definition:

Definition 1: A random function M gives x-differential privacy if for all data sets D1 and D2 differing by at most one element, and for all S such that S ⊆ range (M) [18,19]

$$Pr(M(D1) \in S) \leq exp(x) \times Pr(M(D2) \in S) \quad (1)$$

By applying the DP model, users are allowed to interact with the database only through statistical queries. Other privacy methods such as homomorphic encryption generally release another version of the original database, which provides less efficiency than the DP model. Additionally, DP can prevent risks caused by encryption key leaks. Moreover, it can handle complex and various real data better and blocks more kinds of attacks.

Random noise whose amplitude is chosen according to the L1 sensitivity is added to each query result to reach the DP when the result is numeric. The L1 sensitivity, which is shown in the following definition, is the most significant change that a single participant can have on the output of the query function.

Definition 2: For $f: D \rightarrow Rd$, the L1 sensitivity of f is [18,19]:

$\Delta f = max_{(D1;\, D2)} || f(D1) - f(D2) ||$ *(2)*

for all D1, D2 differing in at most one element.

2. K-Means algorithms

K-means is an unsupervised non-hierarchical clustering algorithm. It allows the observations of the data set to be grouped into K distinct clusters. Thus, similar data will be found in the same cluster. Moreover, an observation can only be found in one cluster at a time (exclusivity of membership). The same observation cannot therefore belong to two different clusters [20,21]. To be able to group a dataset into K distinct clusters, the K-Means algorithm needs a way to compare the degree of similarity between the different observations. Thus, two data that resemble each other will have a reduced dissimilarity distance, while two different objects will have a greater separation distance.

The K-means method is quite simple, starting with selecting the number of clusters as many K chunks, then K chunks of data are randomly taken from the dataset as the centroid to represent a cluster. All data is then calculated at the distance from the centroid and each data will be a member of a cluster represented by a centroid that has the closest distance to the data. Finally, the re-calculation of the centroid value obtained from the average value of each cluster [20,21]. The steps of the K-Means algorithm are as follows [20]:

- ✓ Determine the value of K. (step a)
- ✓ Take random K from the dataset as the centroid. (step b)
- ✓ Calculate the distance of each datum with each centroid (step c)

Definition 3: Calculate the Euclidean distance between two points in space [20]:

$(xi,yj) = \sqrt{((xi_1 - yj_1)^2 + (xi_2 - yj_2)^2 + \cdots + (xi_n - yj_m)^2)}$ (3)

Where (xi, yj): distance between data x and data y; xi: the X data and yj: the Y data.

Definition 4: The average distance between data objects:

$MeanDist = \frac{1}{C_n^2} \sum d(x_i, y_j)$ (4)

Where **n** is the number of data objects in the data set, C_n^2 is the number of couples taken from n data points.

- ✓ Determine the cluster membership of each data. The data becomes a member of the cluster with the nearest centroid distance. (step d)
- ✓ Update the centroid point value. (step e)

$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \boldsymbol{x_i}$ (5)

μk: center of gravity of the cluster K

Nk: the amount of data in cluster K

Xi: Ith datum in cluster K

- ✓ Repeat (step a) to (step e) until there are no more cluster member changes.

3. Model construction

To build our model, we used the google Collaboration platform. To find the optimal number of segments, we use the Elbow method [22]. The *KElbowVisualizer* method implements this result. To create our clusters, we used the K-Means algorithm [20, 22]. The Machine Learning

library Scikit-learn offers us a dedicated function Kmeans. The Figure 3 best describes this research:

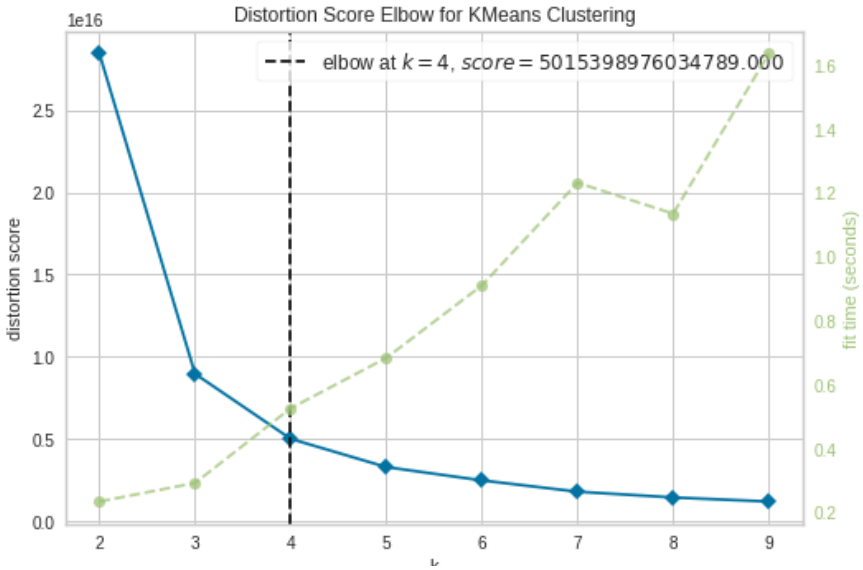


Figure 3: Search for the k-optimal

Figure 3 shows that our data presents four possible segments of users according to the previous criteria. The four cluster centers in Figure 3 are summarized in the following table:

Table 1: Cluster centers of our Model

Classes	Y0	Y1	Y2	Y3
Cluster 0	1.47192515e+05	4.97017478e+06	3.10440651e-02	4.69179039e+01
Cluster 1	8.03080536e+03	2.54517958e+04	4.16832076e-02	1.56269778e+01
Cluster 2	4.72534004e+06	3.77110085e+05	5.39877301e-02	3.54822086e+01
Cluster 3	1.09816685e+06	2.62897327e+06	2.26586103e-02	6.64290030e+01

RESULTS AND DISCUSSIONS

Note that the results obtained make it possible to list the users suspected of fraud. To confirm this, you should have additional information such as the user class, the services subscribed and many others that we coming directly from the operator.

Figure 4 gives us the cases of users suspected of fraud. For representation purposes, our results are presented in a plan. Users suspected of fraud are those whose representation is very far from the centers of the clusters designated by the **yellow dots**.

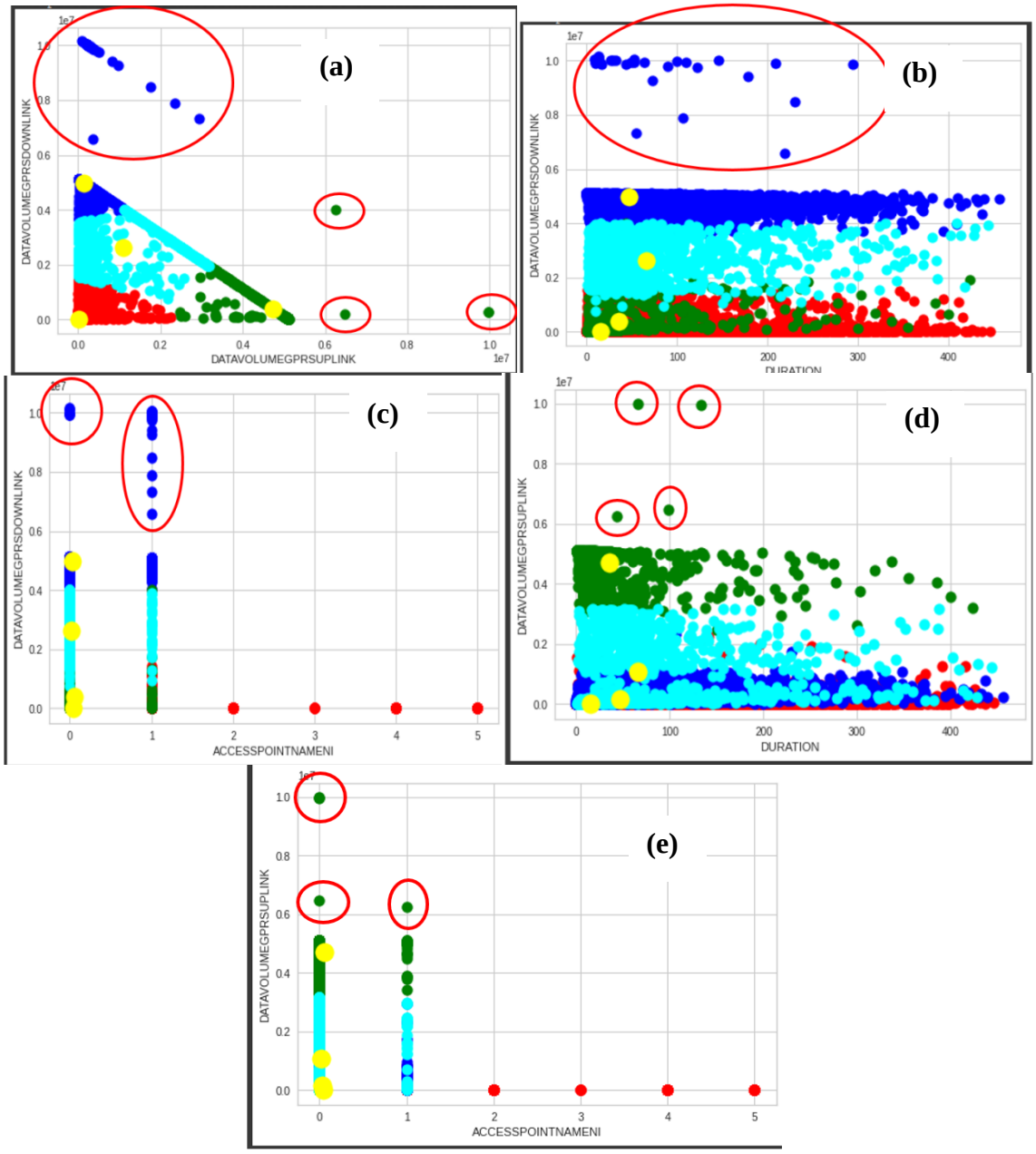


Figure 4: (a) Plan volume of data consumed in UPLINK an DOWLINK (b) Plan volume of data consumed in DOWNLINK and duration of the session, (c) Plan volume of data consumed in DOWNLINK and name of the APN used (d) Plan volume of data consumed in UPLINK and duration of the session (e) Plan volume of data consumed in UPLINK and name of the APN used

In order to evaluate our segmentation model, we will proceed by varying our value of **k.** We obtained as optimal value of k, **k=4**. So, we will evaluate the results ***with k=3 and k=7***. Figure 5, 6 and 7 presents these graphical evaluations and Table 2, 3 and 4 Summarize the results

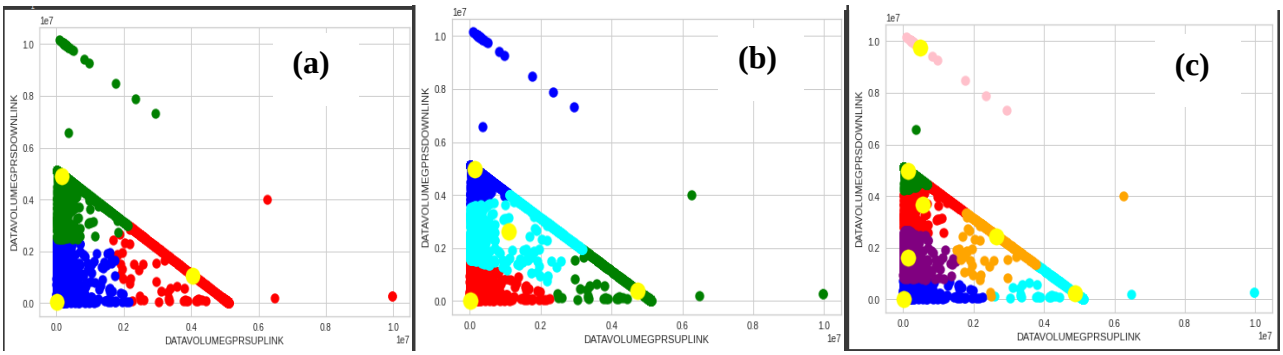


Figure 5: Plan volume of data consumed in UPLINK and DOWNLINK (a) K=3 (b) K=4 (c) K=7

Table2: Summarize of the plan volume of data consumed in UPLINK and DOWNLINK

	K = 3	K = 4	K = 7
Cluster	Confusion between green and red clusters	Each cluster is clearly distributed	Too many clusters that can be grouped together. Example between red and purple

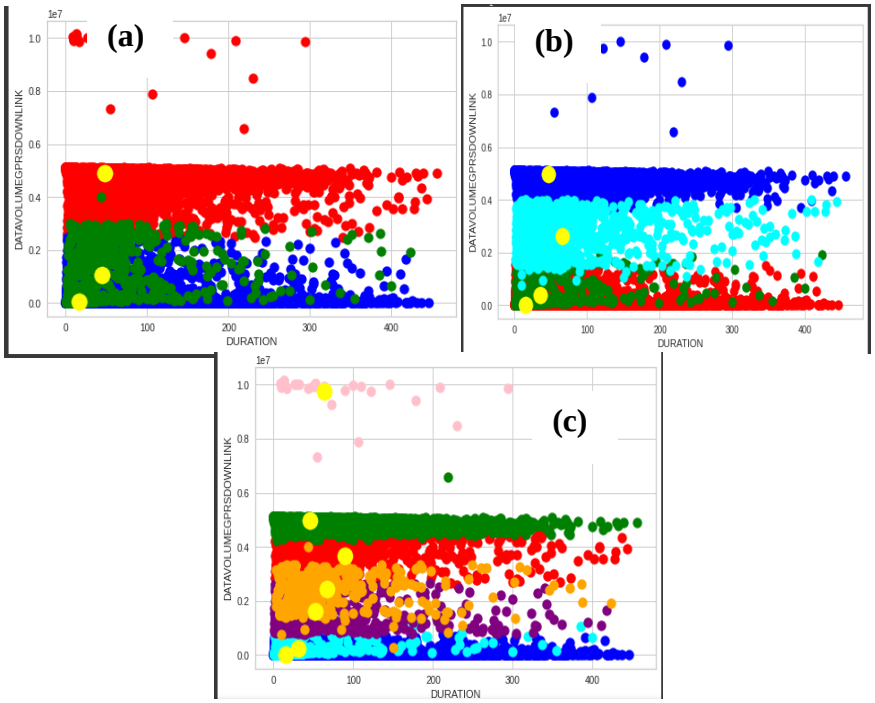


Figure 6: Plan volume of data consumed in DOWNLINK and duration of the session (a) K=3 (b) K=4 (c) K=7

Table 3: Summarize of the plan volume of data consumed in DOWNLINK and duration of the session

	K = 3	K = 4	K = 7
Cluster	Bad segmentation between red, green and blue	Misinterpretation between green and red	Misinterpretation between orange, purple, sky blue and blue

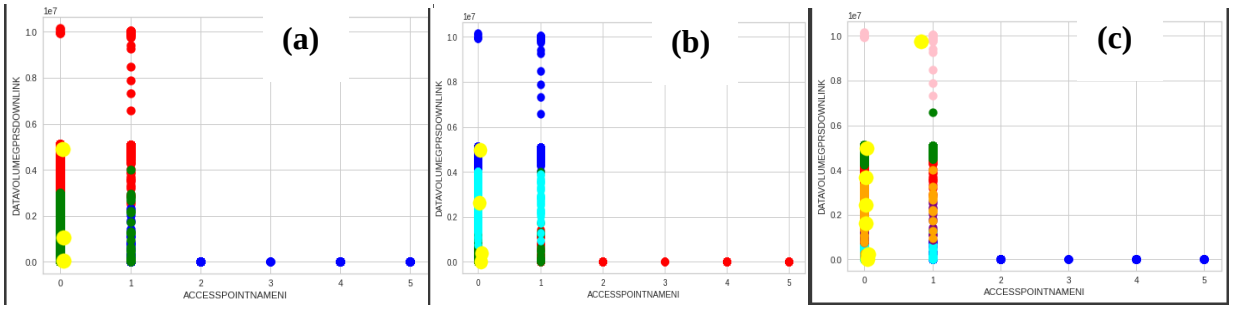


Figure 7: Plan volume of data consumed in DOWNLINK and the name of the APN used (a) K=3 (b) K=4 (c) K=7

Table 4: Summarize of the plan volume of data consumed in DOWNLINK and the name of the APN used

	K = 3	**K = 4**	**K = 7**
Segmentation level	Average	Average	bad

The previous analysis can be summarized in the table below:

	K=3	**K=4**	**K=7**
Assessment of representation 1			
Assessment of representation 2			
Assessment of representation 3			

In view of the previous comparative analysis, our **4-level cluster** presents better results compared to the 3 or 7-level cluster. With a 04-layer segmentation, we clearly distinguish the gap between the different classes of customers.

CONCLUSION

Let us recall in conclusion that the general objective of this work was to define an approach based on "machine learning" for the implementation of a collaborative model for detecting fraud through the analysis of CDRs and traffic to address the problem of fraud in telecommunications networks. Thanks to the data sets obtained on the CDRs of the GGSN and the PCEF, we were able to highlight the four criteria for the behavioral analysis of the users (Volume of data consumed in uplink, Volume of data consumed in Downlink, Name of the APN, Duration of user session). We were thus able to establish several representations indicating the areas belonging to users suspected of fraud. However, we still need subscription information from the subscriber to confirm whether it is a fraudster or not. In addition, we evaluated our segmentation model by comparing it to others. The comparative analysis of the results shows that a four-layer segmentation as we have done offers better results.

Some difficulties have been to confirm whether users suspected of being fraudsters are really so because we don't have labeled data or do not have access to real data on user subscriptions. At the end of our work telecommunications operator can use our model that can interact with their user subscriptions module to detect fraud on data services.

References

[1] Ewieda M, Shaaban EM, Roushdy M (2021) Review of Data Mining Techniques for Detecting Churners in the Telecommunication Industry, Future Computing and Informatics Journal, Vol. 6 : Iss. 1. DOI: http://doi.org/10.54623/fue.fcij.6.1.1

[2] Hadi MS, Lawey AQ, El-Gorashi TEH, JMH (2018) Big data analytics for wireless and wired network design: A survey, Computer Networks 132, 180–199. https://doi.org/10.1016/j.comnet.2018.01.016

[3] S. Gee, Fraud and Fraud Detection: A Data Analytics Approach, Hoboken: John Wiley & Sons, Inc., 2015.

[4] Ouyang Y, Huet A, Shim JP and Hu M (2016) Latent Clustering Models for Outlier Identification in Telecom Data, Mobile Information Systems Vol 2016, Article ID 1542540, 11 p. http://dx.doi.org/10.1155/2016/1542540

[5] Rebahi Y, Nassar M, Magedanz T, Festor O (2011) A survey on fraud and service misuse in voice over IP (VoIP) networks, Information Security Technical report 16, 12-19. doi:10.1016/j.istr.2010.10.012

[6] Ahmed M, Mahmood AN, Hu J (2016) A survey of network anomaly detection techniques, Journal of Network and Computer Applications 60 (2016) 19–31. http://dx.doi.org/10.1016/j.jnca.2015.11.016

[7] Raghavan P, Gayar NE (2019) Fraud Detection using Machine Learning and Deep Learning, IEEE, International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) December 11-12, 2019, Amity University Dubai, UAE.

[8] Alsamhi SH, Almalki FA, Al-Dois H et al (2021) Machine Learning for Smart Environments in B5G Networks: Connectivity and QoS, Computational Intelligence and Neuroscience, Vol 2021, Article ID 6805151, 23 p. https://doi.org/10.1155/2021/6805151

[9] Chouiekha A and EL Hajb E (2018) ConvNets for Fraud Detection analysis, Procedia Computer Science 127, 133–138. Doi 10.1016/j.procs.2018.01.107.

[10] Al-Molhem NR, Rahal Y and Dakkak M (2021) Social network analysis in Telecom data, J. Big Data 6:99. https://doi.org/10.1186/s40537-019-0264-6

[11] Muniyandi AP, Rajeswari R, Rajaramc R (2012) Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm, Procedia Engineering 30,174 – 182. doi:10.1016/j.proeng.2012.01.849

[12] Hoffstadt D, Rathgeb E, Liebig M et al. (2014) A Comprehensive Framework for Detecting and Preventing VoIP Fraud and Misuse, IEEE, DOI: 10.1109/ICCNC.2014.6785441.

[13] Zhao Q, Chen K, Li T, Yang Y, Wang XF (2018) Detecting telecommunication fraud by understanding the contents of a call, Cybersecurity (2018) 1:8. https://doi.org/10.1186/s42400-018-0008-5

[14] Subudhi S and Panigrahi S (2018) A hybrid mobile call fraud detection model using optimized fuzzy C-means clustering and group method of data handling-based network, Vietnam Journal of Computer Science, 5:205–217. https://doi.org/10.1007/s40595-018-0116-x

[15] Sultan K, Ali H, Zhang Z (2018) Call Detail Records Driven Anomaly Detection and Traffic Prediction in Mobile Cellular Networks arXiv:1807.11545v1

[16] Gbadoubissa JEZ, Ari AAA and Gueroui AM (2020), EffiCient k-means based clustering scheme for mobile networks cell sites management, Journal of King Saud University-Computer and Information Sciences, 32(9) 1063-1070. Doi:10.1016/j.jksuci.2018.10.015

[17] Trinh DH, Zeydan E, Giupponi L, Dini P(2019) Detecting Mobile Traffic Anomalies Through Physical Control Channel Fingerprinting: A Deep Semi-Supervised Approach, IEEE Access 7, 152187-152201. DOI: 10.1109/ACCESS.2019.2947742

[18] Dalenius T (1977) Towards a methodology for statistical disclosure control, Statistik Tidskrift, vol. 15, 429-444.

[19] Dwork C (2008) Differential privacy: A survey of results, in Proc. Int. Conf. Theory Appl. Models Comput: Springer. 1-19.

[20] Kouanou AT, Tchiotsop D, Tchinda R, Tchapga TC, Kengnou TAN, Kengne R (2018) "A machine learning algorithm for biomedical image compression using orthogonal transforms", MECS Press I.J Image, Graphics and Signal Processing, 10 (11), 38-53.

[21] Kouanou AT, Tchiotsop D, Fozin Fonzin T, Bayangmbe Mounmo, Tchinda R (2019) "Real-Time Image Compression System Using an Embedded Board", Science Journal of Circuits, Systems and Signal Processing, 7(4), 81-86. doi: 10.11648/j.cssp.20180704.11

[22] Clayman CL, Srinivasana SM, Sangwana RS (2020) K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes, Procedia Computer Science 168, 97–104. Doi: 10.1016/j.procs.2020.02.265