Application of A Multi-Layered Optimal Classifier for Telecommunication Fraud Prediction System

[1]Maureen Akazue, [2]Ifeoma Nonum, [3]Edith Omede and [4]Abel Edje,
[1234]Department of Computer Science, Delta State University, Abraka, Delta State Nigeria

Abstract: This article focuses on the development of a multi-layered optimal classifier to analyze call data records and predict telecommunication fraud. The multi-layered optimal classifier comprises models such as the non-homogeneous poison process (NHPP), Naïve Bayes model, multivariate data sample analysis model, Linear Discriminant Analysis (LDA) leveraging Fisher's model, and Fuzzy function classifier. NHPP is used to estimate the individual subscriber's probabilities of fraud concerning their average number of calls per day and the duration of the calls and this is known as the prior probabilities. The Naïve Bayes model is used to generate the posterior or current probabilities of fraud from the prior probabilities. The multivariate data sample analysis generates the inverse of a variance-covariance pooled dispersion matrices (S-1), which is a component of Fisher's model while Fisher's itself establishes a critical value of posterior probability above which a subscriber is considered to be likely fraudulent. The Fuzzy variable function scans through the call content or conversation of likely fraudulent subscribers for at least a fraud keyword and generates a fuzzy variable zero "0" if it exists, else a fuzzy variable one "1". The Fuzzy function model yields a posterior probability of 1 when the fuzzy variable is zero but the posterior probability remains unchanged if the fuzzy variable is 1.

Keywords: *Classifier; Fraud; Telecommunication; Fuzzy Logic*

I. INTRODUCTION

In today's world especially in Nigeria, telecommunications network management requires extremely rapid decision-making methods that are data mining driven. The Association of Certified Fraud Examiners outlined fraud as "the use of one's occupation for personal enrichment through the deliberate misuse or application of the organization's resources or assets (ACFE, 2019; Ojugo, Akazue, Ejeh.,Odiakaose, and Emordi, May 2023;). Within the technological systems, dishonorable actions have happened in several areas of everyday life like telecommunication networks, mobile communications, online banking, and E-commerce (Akazue, Asuai, Edje, Omede, andUfiofio. July 2023;Akazue and Augusta, 2015;Okofu, Akazue, Ajenaghughrure, and Efozia, 2018). There is exponential growth in fraud as modern technology is growing fast, leading to substantial losses to the companies (Akazue, Onovughe, Omede, and Hampo, 2023;). There is a need to explore more on the fraud detection issue. It involves distinguishing fraud as early as possible once it is been detected.

Fraud is exposed to anomalies in data and patterns. It is quite hard to be certain about the legitimacy of and intention behind an application or transaction. In a real-time network, the volume of collected datasets creates a challenge for analysis, methods, and tools supporting network management tasks (Mwanje, Decarreau, Mannweiler, Naseer-Ul-Islam and Schmelz, 2016; Ojugo and Ekurume, 2021; Okofu, 2018; Okofu, Anazia, Akazue.,Ogeh, Ajenaghughrure, April 2023). For example, how to recognize and identify sudden fraud behavior problems or issues that could prevent large amounts of customer traffic, and how to find network regions and elements that require optimization. These issues are found in routine network management processes.

In the telecommunications industry, fraud continues to affect profitability as the problem results mainly in damages in the financial field since fraudsters are currently "leaching" the revenue of the operators who offer these types of services (Berson, Ngai, and Chau, 2015; Ojie, Akazue, and Imianvan, 2023). The main definition of telecommunication fraud corresponds to the abusive usage of an operator infrastructure, this means, a third party is using the resources of a carrier (telecommunications company) without the intention of paying for them. Other aspects of the problem that cause a lower revenue, are fraud methods that use the identity of legitimate clients to commit fraud, resulting in those clients being framed for a fraud attack that they never committed. This will result in client's loss of confidence in their carrier, giving them reasons to leave. Besides being victims of fraud, some clients do not want to have any business to do with a carrier that has been a victim of fraud attacks. Since the offering of the same services are available from multiple carriers, the client can always switch very easily between them (Berson et al., 2015). Table 1.1 shows the dimension of the "financial hole" generated by fraud in the telecommunications industry and the impact of fraud on the annual revenue of the operators.

This area appealing to some fraudsters. One is the difficulty in the location tracking process of the fraudster; this is a very expensive process and requires a lot of time, which makes it impractical to track a large number of individuals. Another reason is the technological requirements to commit fraud in these systems. A fraudster does not require particularly sophisticated equipment to practice fraud in these systems (Bhattacharya, 2011). All these pieces of evidence generate the need to detect fraudsters most effectivelyto avoid future damage to the telecommunications industry. Data mining techniques are suggested as the most valuable solution since theyallowfor identifying fraudulent activity with a certain degree of confidence. It also works especially well in large amounts of data, a characteristic of the data generated by telecommunications companies (Akazue and Ajenaghughrure, 2016). To detect real-life telecommunication fraud from call data records of customers captured over the network, a sequence of multi-layered fraud prediction models can be applied for optimum results. This calls for the use of the Non-Homogenous Poisson Process (NHPP) to estimate the prior probabilities of the customers' behavior. By developing the Bayesian Statistics, i.e., Naive Bayesian algorithm, posterior probability can be ascertained. From the posterior probabilities / current probabilities of fraud, data samples for multivariate analysis are obtained. Thus, by applying multivariate data sample analysis methodology and developing a linear discriminant analysis for classifying the posterior probabilities of subscriber fraud behavior, a critical value (benchmark) for

Table 1.1: Global Telecommunications Industry Annual Lost Revenue (CFCA Fraud Loss Survey, 2021)

REVENUES	2008	2011	2013	2015	2017	2019	2021
Estimated Global Revenue (USD)	1.7 trillion	2.1 trillion	2.215 trillion	2.254 trillion	2.299 trillion	1.625 trillion	1.8 trillion
Lost Revenue to Fraud (USD)	60.1 billion	40.1 billion	46.3 billion	38.1 billion	29.2 billion	28.3 billion	39.87 billion
% Loss	3.54%	1.88%	2.09%	1.69%	1.27%	1.74%	2.22%

subscriber behavioral classification is obtained. An enhanced optimal classifier can be achieved by exploring the fuzzy variable function in analyzing the call content/conversation to detect fraud keywords and classify customers as regular or fraudulent callers (Akazue, Ojeme, and Anidibia, 2014).

Telecommunication service providers suffer high revenue losses due to fraud. The negative effect of fraud on business success is high especially in this era of stiff competition when robust business intelligence in the mobile telecommunication market is pertinent (Efozia, Anigbogu, and Akazue, 2019). The losses range from the cost of convincing a new customer to use the provider's services to the cost of retaining existing customers. Subscription fraud, social engineering, and even financial fraud are prevalent among many other kinds of fraud carried out on the telecommunication system. Apart from the threat to business success, fraud also poses existential threats like heart attack, depression,etc. to victims. Although efforts have been made by researchers to mitigate telecommunication fraud, these efforts have not yielded significant results as the rate of fraud and its effect is still on the increase.

II. REVIEW OF RELATED WORKS

Similar works abound in the field of artificial intelligence and data mining. Ojugo and Eboka, (2020) developed a magnetic algorithm for short messaging spam filters using text normalization and a semantic approach, while Akazue and Ojeme (2014) built a data mining system for phone businesses. The major application of these varying data mining concepts includes a survey of e-commerce transactional fraud (Akazue, 2015), protection of e-commercetransactions by identifying fake online stores (Akazue, Aghaulor and Ajenaghughrure, 2015), and real-time big data sentiment analysis (Zaki, Hashim and Mohialden, 2020). Other areas of mobile phone communication fraud explored in literature such as using smartphones for phishing and social engineering among Nigerian undergraduate students and mobile phone clients etc are all based on data mining model for fraud (Akazue, Ojugo, Yoro, Malasowe, and Nwankwo 2022); (Eboka and Ojugo 2014)

A. Telecommunication Fraud Dimensions

According to Babaei, Chen, and Maul (2020), fraud was highlighted as a serious issue in most telecommunication systems as it leads to financial losses as well as loss of customers. Telecommunication fraud represents the abusive usage of an operator's infrastructure. The carrier and the client are the major victims. Apart from financial reasons, other motivations such as political motivations, personal achievements, and self-preservation make the criminals motivated to commit the attacks (Becker et al, 2009). Some identified fraud dimensions include superimposed fraud, subscription fraud, technical fraud, internal fraud, network anomaly-based intrusion, social engineering, and hybrid shoer message service spamming (João and Sousa, 2014; Kimmo, 2009; Ojugo, Oyemade, and Ekurume, 2021). Figure 2.1 highlights the fraud methods and incidences in the year 2019 worldwide (CFCA Fraud Survey, 2019).

B. Analysis of the Existing System

The existing system is an optimal computation model developed by Amujiet al.(2019). They analyzed synthetic simulated data generated from Minitab software. They simulated a sample of eighty (80) subscribers: their number of calls and the duration of the calls and categorized it into four sub-samples with a sample size of twenty (20) each. They obtained the prior and posterior probabilities of the groups and these posterior probability distributions were grouped into two sample multivariate data with two variates each. They developed the linear classifier that discriminates between genuine subscribers and fraudulent subscribers. The optimal classifier (β A B+) has a posterior probability of 0.7368, and they classified the subscribers based on this optimal point. Their work focused on domestic subscribers and the parameters of interest were the number of calls per hour and the duration of the calls.

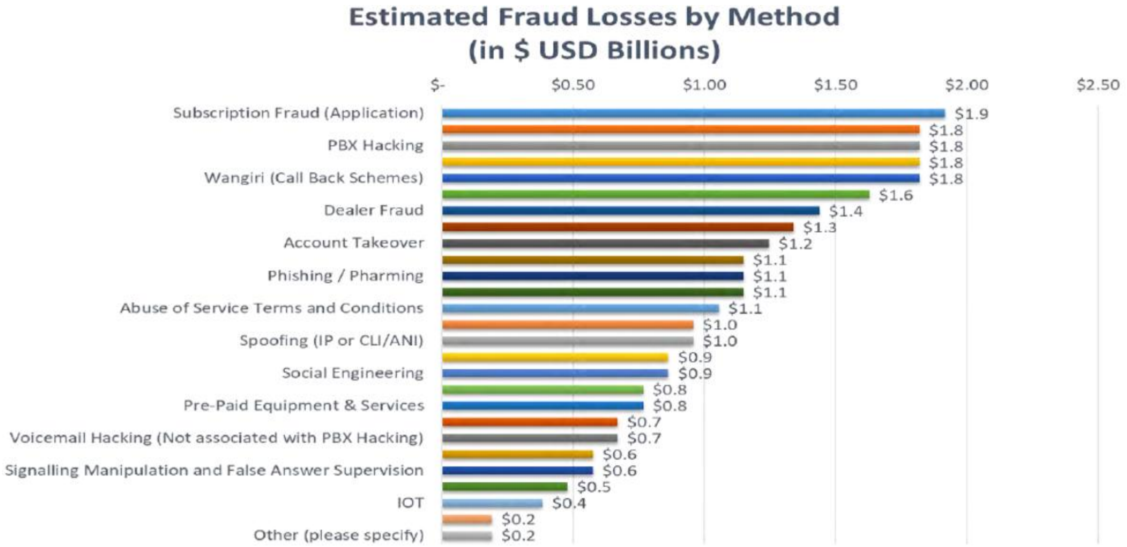

Figure 2.1: World Wide Telecommunications Fraud Methods Statistics, 2019.

C. Weakness of the Existing System in Relation the Proposed System

The work classified subscribers as fraudulent or non-fraudulent only based on the posterior probability values which is a function of call frequency and duration. This classification benchmark will not give high accuracy because a genuine customer could have a very high posterior probability due to hyperactivity arising from legitimate business opportunities. Therefore, the need for a fuzzy function classifier which is dependent on the existence of a fraud keyword in the subscriber's call conversation cannot be overemphasized. The proposed model includes a fuzzy function as a part of the multi-layered architectural model. The multi-layered architectural model approach solves the problems of inaccuracy arising from the computation of unstructured/imbalanced fraud datasets by reducing the number of customers classified as false positive, and thuds, improving the accuracy of the classifier.

III. MODEL DEVELOPMENT

This session focuses on the development of the various models that can be applied in building a computational model for telecom fraud detection.

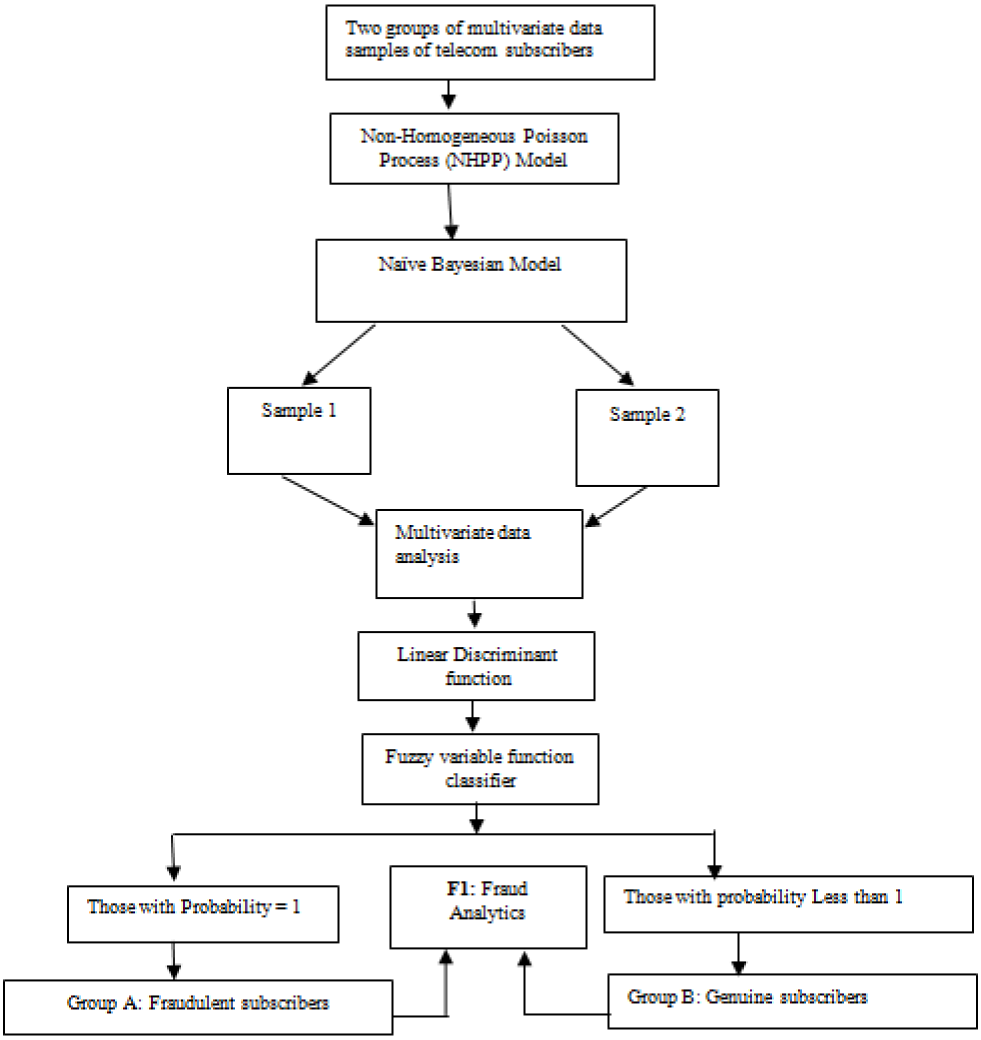
Two groups of multivariate data samples of telecom subscribers
Non-Homogeneous Poisson Process (NHPP) Model
Naïve Bayesian Model
Sample 1
Sample 2
Multivariate data analysis
Linear Discriminant function
Fuzzy variable function classifier
Those with Probability = 1
F1: Fraud Analytics
Those with probability Less than 1
Group A: Fraudulent subscribers
Group B: Genuine subscribers

Fig 3.1: Tree Diagram for a Multi-Layered Optimal Classifier for Telecommunication Fraud prediction

A. Non-Homogeneous Poison Process

This is a subset of Poisson distribution which provides a realistic model for random phenomena. Since the value of Poisson random variables are the non-negative integers, a random phenomenon for which a count is of interest is a candidate for modeling by assuming a Poisson distribution.

Such a count included the number of telephone calls per unit of time coming into the switchboard of a large business (service providers). If certain assumptions regarding the phenomenon under observation are satisfied, the Poisson model is the correct model (Mostafaet al., 2018).

Hence, if {X_n,n≥0} is a sequence of independent identically distributed exp(λ) random variables, the counting process {N(t),n≥0} is called a Poisson process with parameter λ and it is denoted by PP(λ). Thus, the first event counted by N(t) takes place after an exponential amount of time with parameter λ. The rest of the inter-event time are independent identically distributed exponential with parameter λ. NHPP(λt) can be thought of as a process that counts events that occur in a non-uniform fashion (Barlow and Proschan, 1975; Field and Zidek, 1995), hence this work established critical assumptions for NHPP(λt).

Model Assumptions

The accounting process $\{N(t), n \geq 0\}$ is called a non-homogeneous Poisson Process (NHPP(λt)) with a rate function $\{\lambda(t), t \geq 0\}$ if

The number of events at time zero is equal to zero

i. The number of events in non-overlapping time interval are independent

ii. $o(h) \rightarrow$ some function of smaller order than which satisfies the condition $\lim_{h\to 0}\frac{o(h)}{h}=0$

iii. The probability that exactly one event will occur in a small interval of time t + h approximately equal to $\lambda(t).h$ or

iv. The probability that no event occurs in the time interval t + h is given by $1-\lambda(t).h+0(h)$.

v. The probability that more than one event will occur in a small interval of time t + h is negligible

vi. The events must occur at random.

Now, that it is feasible to write *{N(t), t ≥ 0)}* ~ *NHPP (λ(.))*. This denotes that *{N(t), t ≥ 0)* is a non-homogeneous Poisson process with rate function *λ(.)*. When $\lambda(t)=\lambda$ for all t ≥ 0, then NHPP becomes an HPP. Thus, NHPP is a generalization of HPP. In both HPP and NHPP, events take place one at a time (Field and Zidek, 1995).Hence, the predictive probability models for this work are non-homogenous Poisson processes. This is then given in eqn. 3.5, while the Bayesian statistics model is given in equation (3.6)

$$P_n(t)=\exp\{-\lambda t\}\left[\frac{\{\lambda t\}^n}{n!}\right] \quad ;n=0,\dots \quad (3.1)$$

Where $P_n(t)$ = the probability of *n* number of calls at a given time *(t)*, λ is the parameter (intensity) of the model and *t* is time in minutes. Equation (3.1) model was used to determine the prior distribution.

B. Naïve Bayes Model

Suppose the events A_1, A_2, . . . , A_n , representing different callers, form a sample space Ω, that is, the events A_i are mutually exclusive but collectively exhaustive and their union is Ω. Let B be any other event, say a fraudulent subscriber, defined on a sample space Ω.

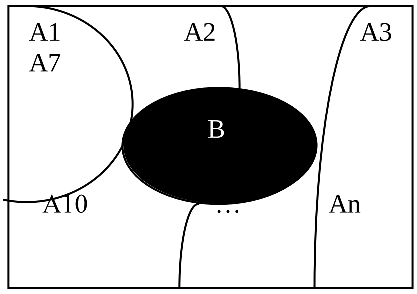

Figure 3.2: Venn Diagram Characterising Mobile Telecommunication Ecosystem

From Figure 3.2, let *B* given as same as:

$$B=\Omega\cap B$$

$$\Omega=A_1\cup A_2\cup A_3\cup\ldots\cup A_n$$

$$B=(A_1\cup A_2\cup A_3\cup\ldots\cup A_n)\cap B$$

Now the joint or total probability (likelihood function) *P(B)* can be determined from eqn. 3.21

$$B=(A_1\cap B)\cup(A_2\cap B)\cup(A_3\cap B)\cup\ldots\cup(A_n\cap B)$$

$$\equiv\bigcup_{i=1}^{n}(A_i\cap B)$$

$$P(B)=P(A_1\cap B)+P(A_2\cap B)+P(A_3\cap B)+\ldots$$
$$+P(A_n\cap B)$$

$$P(B)=P(A_1)P(B/A_1)+P(A_2)P(B/A_2)+\ldots$$
$$+P(A_n)P(B/A_n)$$

$$P(B)=\sum_{i=1}^{n}P(A_i)P(B/A_i)$$

$$P(A_i/B)=\frac{P(A_i\cap B)}{P(B)}=\frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n}P(A_i)P(B/A_i)}$$

Equation (3.5) was used to simply validate Figure 3.2 for Proof of Naive Bayes (Ω)

Similarly, if the events θ_1, θ_2, . . . , θ_n form a sample space Ω, i.e., the events Θ_i are mutually exclusive but collectively exhaustive and their union is Ω. Let *y* be any other event defined on a sample space Ω. Then, the Bayes theorem states that

$$P(Y=y_i/\theta)=\frac{P(y/\theta_i).P(\theta_i)}{\sum_{i=1}^{n}P(y/\theta_i).P(\theta_i)}$$

$$P(y)=\sum_{i=1}^{n}P(\theta_i)P(y/\theta_i)$$

$$P(Y=y/\theta_i)=\frac{P(y\cap\theta_i)}{P(y)}=\frac{P(y/\theta_i).P(\theta_i)}{\sum_{i=1}^{n}P(y/\theta_i).P(\theta_i)}$$

Where $P(Y=y/\theta)$ = the conditional probability that the random variable Y assumes a specific value y given that its prior probability was θ. Note that $\theta=\lambda$ is now a random variable. = the likelihood function of the distribution.

Equation (3.6) is now the estimated Bayesian statistics model used to determine the posterior probability. Hence, the predictive probability data mining model is the model in equation (3.6). The model is the Bayesian statistical model.

C. Multivariate Data Analytic Models

This model is used to analyze the posterior probabilities derived from the Naïve Bayes model where x_1, x_2, x_3, x_{n+1} are the random call data variables or average call duration (ACD) while $\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$, $\overline{x_{n+1}}$ represents the mean random call data variables/ACD.

In the case of fraud analysis, pooled sample dispersion matrix, dispersion matrix, and inverse of the dispersion matrix were determined for the developed computational model. The dispersion (variance-covariance) matrix with three variants is given by equation (3.7).

$$(n_i-1)S_i^{\,2}=\begin{bmatrix}\sum x_1^{\,2}-n\bar{x}_1^{\,2} & \sum x_1x_2-n\bar{x}_1\bar{x}_2 & \sum x_1x_3-n\bar{x}_1\bar{x}_3\\ \sum x_1x_2-n\bar{x}_1\bar{x}_2 & \sum x_2^{\,2}-n\bar{x}_2^{\,2} & \sum x_2x_3-n\bar{x}_2\bar{x}_3\\ \sum x_1x_3-n\bar{x}_1\bar{x}_3 & \sum x_2x_3-n\bar{x}_2\bar{x}_3 & \sum x_3^{\,2}-n\bar{x}_3^{\,2}\end{bmatrix}$$

where n_1 and n_2 respectively stand for the first and second samples. The variances (a measure of the width of distribution for the call samples) are

$$X_{11}=\sum_{1}^{30}X_1^{\,2}-n\overline{X}_1^{\,2};\; X_{22}=\sum_{1}^{30}X_2^{\,2}-n\overline{X}_2^{\,2};$$

$$X_{33}=\sum_{1}^{30}X_3^{\,2}-n\overline{X}_3^{\,2}$$

The covariance is

$$X_{12}=\sum_{1}^{30}X_1X_2-n\overline{X}_1\overline{X}_2;\; X_{13}=\sum_{1}^{30}X_1X_3-n\overline{X}_1\overline{X}_3;$$

$$X_{23}=\sum_{1}^{30}X_2X_3-n\overline{X}_2\overline{X}_3; \qquad (3.8)$$

The model representation in equation (3.8) is a symmetric matrix, where the diagonal elements are the variances which cannot be negative. The upper and lower entries are the covariance

D. Linear Discriminant Function

The linear discriminant analysis is introduced to discriminate between the normal subscribers and those who are fraudulent. The main idea of discriminant analysis is a search for the differences in two or more groups that consist of multivariate measurements. One (or more) linear function(s) which maximally differentiate(s) between these groups are constructed. These functions are then used to classify new members of similar groups into the appropriate group they belong and differentiate them from the group they do not belong. The main facts that do the differentiation are contained in the pooled Dispersion matrix of the two groups (Onyeagu, 2003). Now, the Classification into one of two populations with known probability distributions is presented as:

Let the probability that observation comes from subscriber population K_1 be q_1 and from K_2 be q_2. Let the probability distribution function (pdf) of $K_1 = P_1(x)$ and $K_2 = P_2(x)$.LetR_1 and R_2 be the regions of classification corresponding to populationsK_1 and K_2. Let R_1 *and*R_2be defined as

$$R_1:\frac{P_1(x)}{P_2(x)}\geq\frac{C(1/2)q_2}{C(2/1)q_1},$$ i.e.,the Cost of classifying population 1 into subscriber population 2, (3.9)

$$R_2:\frac{P_1(x)}{P_2(x)}<\frac{C(1/2)q_2}{C(2/1)q_1}$$

The above shows that R_1 and $R_2$are Bayes procedures, that is, R_1 and R_2 minimize the expected loss from costs of misclassification. The classification into one of two known multivariate normal populations is given as:

Let $K_1\sim N_p\left(\mu^{(1)},V\right)$ and $K_2\sim N_p\left(\mu^{(2)},V\right)$

$$P_i(x)=(2\pi)^{-p/2}|V|^{-1/2}\exp\left[-1/2\left(x-\mu^{(1)}\right)^TV^{-1}\left(x-\mu^{(1)}\right)\right];\, i=1(3.10)$$

$$\frac{P_1(x)}{P_2(x)}=\frac{(2\pi)^{-p/2}|V|^{-1/2}\exp\left[-1/2\left(x-\mu^{(1)}\right)^TV^{-1}\left(x-\mu^{(1)}\right)\right]}{(2\pi)^{-p/2}|V|^{-1/2}\exp\left[-1/2\left(x-\mu^{(2)}\right)^TV^{-1}\left(x-\mu^{(2)}\right)\right]} \qquad (3.11)$$

$$=\exp\left[-1/2\left(x-\mu^{(1)}\right)^TV^{-1}\left(x-\mu^{(1)}\right)-\left(x-\mu^{(2)}\right)^TV^{-1}\left(x-\mu^{(2)}\right)\right]$$

Then the optimal determination of the region x for which an individual is classified to K$_1$ is

$$W=\left\{x:\frac{P_1(x)}{P_2(x)}\geq K\right\} \qquad (3.12)$$

Where $K=\dfrac{C_{1,2}q_2}{C_{2,1}q_1}$

By takingthe natural log of both sides, this gives equation (3.12)

$$W=\left\{x:\log\frac{P_1(x)}{P_2(x)}\geq\log K\right\} \qquad (3.13)$$

$$W=\left[x:-1/2\left(x-\mu^{(1)}\right)^TV^{-1}\left(x-\mu^{(1)}\right)-\left(x-\mu^{(2)}\right)^TV^{-1}\left(x-\mu^{(2)}\right)\right]\geq\log K \qquad (3.14)$$

The left-hand side (LHS) of equation (3.13) inequality can be expanded and rearranged to give

$$W=-1/2\left[x^TV^{-1}x-x^TV^{-1}\mu^{(1)}-\mu^{(1)T}V^{-1}x+\mu^{(1)T}V^{-1}\mu^{(1)}-(x^TV^{-1}x-x^TV^{-1}\mu^{(2)}-\mu^{(2)T}V^{-1}x+\mu^{(2)T}V^{-1}\mu^{(2)})\right] \qquad (3.15)$$

$$W=-1/2\left[-2x^TV^{-1}\mu^{(1)}+2x^TV^{-1}\mu^{(2)}+\mu^{(1)T}V^{-1}\mu^{(1)}-\mu^{(2)T}V^{-1}\mu^{(2)}\right] \qquad (3.16)$$

$$W=-1/2\left[-2x^TV^{-1}(\mu^{(1)}-\mu^{(2)})+(\mu^{(1)}-\mu^{(2)})^TV^{-1}(\mu^{(1)}+\mu^{(2)})\right] \qquad (3.17)$$

$$W=x^TV^{-1}(\mu^{(1)}-\mu^{(2)})-1/2(\mu^{(1)}+\mu^{(2)})^TV^{-1}(\mu^{(1)}-\mu^{(2)}) \qquad (3.18)$$

From equation (3.18), the first component of the function [$x^TV^{-1}(\mu^{(1)}-\mu^{(2)})$] is Fisher's Linear Discriminant function, where the population dispersion matrix V^{-1}is estimated by sample dispersion matrix $S^{-1.}$ Hence, the discriminant function employed in this work is called Fisher's linear discriminant function given in equations (3.19),

$$W=X^TS^{-1}\left(\overline{X}^{(1)}-\overline{X}^{(2)}\right) \qquad (3.19)$$

Where $X^T=\left(X_1,\ldots,X_p\right)$; S^{-1} is the inverse of the dispersion (var) matrix and $\left(\overline{X}^{(1)}-\overline{X}^{(2)}\right)$ is the difference in the mean vectors between two multivariate samples and W is the linear discriminant function.

E. Fuzzy Variable Function Classifier

The last part of the analysis is the introduction of fuzzy function $(\sigma_i), 0\leq i\leq 1$, into the model for classification purposes. The fuzzy function takes the value zero *(0)* where it is established that the subscriber uses any of the sensitive words from the fraud sensitivity list, indicating fraudulent practices (e.g.*"send your ATM pin","free call", "browsing cheat", "BVN"*), and one (1) otherwise

$$\sigma_i = \begin{cases} 0 & \text{if sensitive words are used} \\ 1 & \text{otherwise} \end{cases} \qquad (3.20)$$

The probability classification model for fraud detection among subscribers in any telecommunication system is

$$P(Y_i) = \left[\frac{P(y/\theta_i).P(\theta_i)}{\sum_{i=1}^{n} P(y/\theta_i).P(\theta_i)} \right]^{\sigma_i} \qquad (3.21)$$

The model in equation (3.21) is the probability classifier model, which has not been developed or used before. The classification rule is now: classify the subscribers with a probability of 1 as fraudulent subscribers and those with a probability less than one as genuine subscribers. When the subscribers have probability one, it implies that for sure, the subscribers are fraudulent customers.

CONCLUSION

A multi-layered optimal classifier for telecommunication fraud prediction leveraging probability models, linear discriminant, and fuzzy functions is used to analyze subscribers' call detail records and classify them into fraudulent and non-fraudulent subscribers. The multi-layered optimal classifier model is applied in a complex telecommunication ecosystem in which end users' devices are connected to the base transceiver station and the calls or communication between these subscribers are captured and analyzed for fraud behaviour with the help of the prediction model running in the mobile switching layer. It uses the non-homogeneous Poisson Process model (NHPP) to estimate the individual subscriber's probability of fraud, judging from the subscriber's call to the average number of calls made by all the subscribers in the same cluster. The estimated probability from the NHPP model assumes the random variable in the Naïve Bayes model to calculate the prior and posterior probabilities of fraud of the individual subscribers. A telecommunication fraud classifier is built using Fisher's model to classify subscribers based on their posterior probabilities. Also, a fuzzy function classifier which classifies based on the existence of fraud keywords is developed and this gives a higher classification accuracy. The multi-layered optimal classifier relies on detecting abnormal events. These abnormal events are characterized by relating the events to symptoms associated with fraudulent events in the past. This multi-level architecture enables effective fraud predictions from huge imbalanced data.

The multi-layered optimal classifier architectural model has given better classification results and accuracy when used to implement the classification for fraud prediction of real-life call data records captured from telecommunication networks.

References

[1] Akazue M. I, Ajenaghughrure I. B. (2016), A Survey and Cost Classification of Big Data Analytics and Decision Tools, International Journal of Innovative Research in Computer and Communication 2016

[2] Akazue M. I., Ojugo A. A., Yoro R. E., Malasowe B. O., and Nwankwo O. (2022), "Empirical evidence of phishing menace among undergraduate smartphone users in selected universities in Nigeria," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 28, no. 3, pp. 1756–1765, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1756-1765.

[3] Akazue M.I, Ojeme B, Anidibia D (2014), A Computerised Approach of Statistical Inference, oriental journal of Computer Science and technology 7 (2), 237-243 2014

[4] Akazue Maureen and Ojeme Blessing (2014), "Building Data Mining For Phone Business", Oriental Journal of computer science and Technology: An international open access peer reviewed research journal, vol. 7, no. 03, pp. 316-322, 2014.

[5] Akazue, M . I., Aghaulor, A., and Ajenaghughrure, B. I. (n.d.). Customer's Protection in Ecommerce Transaction Through Identifying Fake Online Stores. In International Conference e-Learning, e-Bus., EIS, and e-Gov. | EEE'15 (pp. 52–54).

[6] Akazue, M. I. (2015), "A Survey of Ecommerce Transaction Fraud Prevention Models." In The Proceedings of the International Conference on Digital Information Processing, Data Mining, and Wireless Communications, Dubai, UAE. 2015.

[7] Akazue, M., and Augusta, A. (2015). Identification of Cloned Payment Page in Ecommerce Transaction. International Management Review, 11(2), 70-76.

[8] Akazue M., Asuai C., Edje A., Omede E., Ufiofio E. (July 2023) CYBERSHIELD: Harnessing ensemble feature selection technique for robust distributed denial of service attacks detection, Kongzhi yu Juece/Control and Decision, Volume 38, Issue 03, July, pp 1211-1224

[9] Akazue M., Onovughe A., Omede E., Hampo J.A.C. (2023) Use of Adaptive Boosting Algorithm to Estimate Users Trust in the Utilization of Virtual Assistant Systems, International Journal of Innovative Science and Research Technology. Vol.8, Issue 1, Pages: 502 – 507. https://www.ijisrt.com/assets/upload/files/IJISRT23JAN727.pdf

[10] Amuji, H.O., Chukwuemeka, E. and Ogbuagu, E.M.(2019) Optimal Classifier for Fraud Detection in Telecommunication Industry. *OpenJournal of Optimization*, **8**, 15-31.

[11] Association of Certified Fraud Examiners (ACFE) (2019). Report to the Nations Global Study on Occupational Fraud and Abuse Government. New York: Association of Certified Fraud Examiners.

[12] Babaei, K., Chen, Z. Y. and Maul, T. (2020). A Study of Fraud Types, Challenges and Detection Approachesin Telecommunication. Journal of Information Systems and Telecommunication, 7(4), pp. 248 – 261.

[13] Barlow, R.E, and Proschan, F. (1975). Statistical Theory of Reliability and Life Testing probability models. Holt, Rinehart and Winston, Inc. USA.

[14] Becker, R.A., Volinsky, C., and Allan R. W. (2009). Fraud Detection in Telecommunications: History and Lessons Learned. Technometrics, 52(1):20–33, February 2010. URL: http://www.tandfonline.com/doi/abs/10.1198/TECH.2009.08136, Doi:10.1198/TECH.2009.08136.

[15] Berson, W.T., Ngai, L. X. and Chau, D.C.K. (2015). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592 – 2602.

[16] Bhattacharya, C. (2011). When customers are members: customer retention in paid membership contexts. Journal of the Academy of Marketing Science, 26, 31-44.Communications Fraud Control Association (CFCA) – Fraud Loss Survey 2021

[17] Eboka A. and Ojugo A. A.(2014), “A social engineering detection model for the mobile smartphone clients,” African Journal of Computing & ICT, vol. 7, no. 3, pp. 91–100, 2014. [14]

[18] Efozia N. F., Anigbogu S. O. and Akazue M. I. (2019), A Review on the Concept of Business Intelligence, International Journal of Trend in Research and Development 6 (5), 160-169, 2019

[19] Field, C. and Zidek, J.V. (1995). Modeling and Analysis of Stochastic Systems. Chapman and Hall, London, UK.

[20] João V., and De Sousa, C. (2014). Telecommunication Fraud Detection Using Data Mining Techniques. M.sc thesis, 2014, Department of Electrical and Computers Engineering, University of Porto, Porto.

[21] Kimmo S. H (2009). Designing an expert system for fraud detection in Private Telecommunication’s networks. Expert Systems with Applications. 36(9), pp.11559– 11569,

[22] Mostafa, H. Anwer, M. and Abdel-Hamid, F.A. (2018). A framework for efficient network anomaly intrusion detection with features selection, 2018 9th International Conference on Information and Communication Systems (ICICS), 3-5 April 2018, DOI: 10.1109/IACS.2018.8355459.

[23] Mwanje, S., Decarreau, G., Mannweiler, C., Naseer-ul-Islam, M., and Schmelz, L.C (2016). “Network management automation in 5G: Challenges and opportunities”, In Proc. 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 4-8 Sept. 2016, Valencia, Spain.

[24] Ojie, D.V, Akazue M, and Imianvan A. (2023). A Framework for Feature Selection using Data Value Metric and Genetic Algorithm, International Journal of Computer Applications, Vol 184, Issue43, p14-21, doi:10.5120/ijca2023922533

[25] Ojugo, A.A., Akazue, M.I., Ejeh, P.O., Odiakaose, C.C., Emordi, F.U. (May, 2023). DeGATraMoNN: Deep Learning Memetic Ensemble to Detect Spam Threats via a Content-Based Processing. Kongzhi yu Juece/Control and Decision, Volume 38, Issue 01, April 2023, ISSN: 1001-0920

[26] Ojugo A. A. and Eboka A. O.(2020), “Memetic algorithm for short messaging service spam filter using text normalization and semantic approach,” International Journal of Informatics and Communication Technology (IJ-ICT), vol. 9, no. 1, pp. 9–18, Apr. 2020, doi: 10.11591/ijict.v9i1.pp9-18. [42]

[27] Ojugo A. A. and Ekurume E. (2021), “Deep learning network anomaly-based intrusion detection ensemble for predictive intelligence to curb malicious connections: An empirical evidence,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 3, pp. 2090–2102, Jun. 2021, doi: 10.30534/ijatcse/2021/851032021.

[28] Ojugo A. A. and Oyemade D. A. (2021), “Boyer Moore string-match framework for a hybrid short message service spam filtering technique,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 3, pp. 519–527, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp519-527

[29] Okofu, S. N. (2018). Users Service Quality Trust Perception of Online Hotel Room Reservation. SAU Journal of Management and Social Sciences,Vol 3, No.1 & 2, pp 1-14

[30] Okofu, S. N., Akazue, M. I., Ajenaghughrure B. I., and Efozia F. N. (2018). An Enhanced Speech-Based Airline Ticket Reservation Data Confirmation Module. Journal of Social and Management Sciences. Vol 13, issue 1, pg. 11-21.

[31] Okofu, S., Anazia E. K., Akazue M., Ogeh C., and Ajenaghughrure, I. B. (April 2023), The Interplay Between Trust In Human-Like Technologies And Integral Emotions: Google Assistant. Kongzhi yu Juece/Control and Decision, ISSN: 1001-0920. Vol 38, Issue 01

[32] Onyeagu, S. I. (2003). First Course in Multivariate Statistical Analysis. Mega Concept Publishers, Awka, Nigeria.

[33] Y Zhang, S Wang, P Phillips, G Ji, Binary PSO with mutation operator for feature selection using decision tree applied to spam detection, Knowledge-Based Systems 64, 22-31

[34] Zaki ND, Hashim NY, Mohialden YM, Mohammed MA, Sutikno T, Ali AH. (2020), A real-time big data sentiment analysis for Iraqi tweets using spark streaming, Bulletin of Electrical Engineering and Informatics 9 (4), 1411-1419