International Journal of Advanced Engineering and Technology
www.allengineeringjournal.com
ISSN: 2456-7655
Received: 01-06-2023, Accepted: 16-06-2023, Published: 01-07-2023
Volume 7, Issue 2, 2023, Page No. 1-4

Enhanced predictive data mining algorithm for fraud detection and churn behaviour modelling in telecommunication systems

Promise Elechi, Iwok Odudu-Abasi Michael
Center for Information and Communication Technology University of Port Harcourt, Nigeria

Abstract
An improved data mining predictive model for fraud detection and churn behavior in a telecommunication network system is presented in this thesis. The large revenue losses suffered by telecom service providers as a result of fraud served as the impetus for this study. Telecommunications fraud detection and prevention includes any strategy or procedure used to minimize illicit activity meant to hurt telecom service providers. The losses include anything from the price of persuading a brand-new client to use the services of the supplier to the price of keeping hold of current clients. The complicated network architecture that categories and retains derived patterns in the cloud back-end for analytics was the subject of the study. To develop an adaptive control strategy for fraud detection, computational analytic modelling was used in conjunction with probabilistic models, a Naive Bayesian model, a linear discriminant function, and neural prediction networks. To take use of call detail records travelling non-homogeneously through the network, simulation and train-classification technique were investigated. Modularization, class containers, and data structures were used to create the Java-SQL Containerization technique. When describing the communication network as having edge devices coupled to the Base transceiver (BT) controllers, the global system architecture for Fraud behavior synthesis was described. The subscriber/sample radial basis neural network function (SRBNF) was constructed from the network using computational neural controller architecture. A predictive probability data mining model was created during the data analysis employing the Non-Homogenous Poisson Process (NHHP (t)) for prior knowledge. The posterior probability was calculated using the Naive Bayes (supervised learning) classification algorithm for low- and high-income subscribers. For a multivariate analysis, a critical threshold discriminant function (CTDF) Value of 0.00229 was achieved. As a result, a pooled sample dispersion matrix was created, as well as inverse dispersion matrices for the neural computational model that was created. The proposed data mining predictive model provided a Mean Square Error (MSE) for the CTDF of 1.7562 in the SRBNF validation. To find the best algorithm or model that produces accurate, dependable outcomes consistently, an examination was conducted. Therefore, three algorithms—Decision Tree (DT), Logistic Regression (LR), and Enhanced Neural Discriminant Analysis (Proposed)—were examined for fraud detection. These yielded, respectively, 14.29%, 30.00%, and 55.71%. The best and most accurate prediction threshold for fraud attrition was thus found to be provided by the suggested method.

Keywords: Data mining, fraud detection, churn behavior, telecommunication network system, computational analytic modelling, probabilistic models, naive bayesian model, linear discriminant function, neural prediction networks, Java-SQL containerization technique

Introduction

In the context of Nigeria's telecommunications industry, the efficient operation of networks relies heavily on data mining and fast decision-making processes. Churn and fraud management in telecommunications networks are based on data extracted from continuous data collection. With the rapid growth of telecommunication networks and the increasing volume of surveying data, churn and fraudulent activities have also increased. The expansion of mobile technology and the development of new end-user services have fueled the growth of data in the industry.

Telecommunication networks rely on data provided by network elements, which generate log records, event notifications, system status updates, and performance data. These data are transmitted to command or control centres, where they are monitored and analysed to identify performance and service quality issues. The sheer amount of data collected in real-time poses a challenge for network management.

Data mining and knowledge discovery have emerged as innovative approaches for analyzing complex data in systems. These methodologies integrate automated reasoning, neural models, ML “machine learning”, statistics, plus online data analysis. telecom systems, known for generating substantial data volumes, were among the pioneers in adopting data mining techniques. Numerous strategies have been devised to combat fraud and reduce customer attrition in telecommunications networks, necessitating a comprehensive understanding of network architecture, communication technologies, consumer behaviour, and user preferences.

Fraud detection and prevention play a crucial role in risk management for telecommunication companies. Failure to incorporate fraud detection and prevention technologies can result in significant financial losses and a decline in the subscriber base. Larger telecommunications companies are particularly vulnerable to fraud due to their ability to resell services and networks to smaller operators. By understanding the types of fraud prevalent in the telecommunications industry, operators can proactively combat these scams and mitigate their impact.

Overall, in Nigeria's telecommunications industry, data mining and advanced analytics techniques are essential for effective network management, fraud detection, and churn

prevention. These technologies help operators make data-driven decisions and mitigate risks associated with fraudulent activities, ultimately improving their financial performance and subscriber satisfaction.

Forms of fraud in the telecommunications industry

Three major target categories can be used to categorize telecommunication frauds:

- fraud against subscribers
- fraud against telecommunication service providers
- general telephone fraud

Let's examine the many telecommunication fraud schemes.

International revenue sharing fraud (IRSF): occurs when fraudsters get into a company's phone system and lease a premium phone number from which they place calls. As a result of the revenue sharing mechanism, which charges commission to the business that rented a premium number from the IPRN (for directing callers to the number) as well as commission to the owners of premium rate numbers so that they can include a portion of the call earnings in their monthly invoicing services or real-time phone crediting programs, the company is forced to pay exorbitant call rates (up to $1 per minute), some of which end up in the hands of fraudsters.

so that they can include a portion of the call earnings in their monthly invoicing services or real-time phone crediting programs.

Fraudsters purchase a local carrier's SIM card and use a SIM box or a GSM gateway to reroute the international calls in order to commit interconnect bypass fraud (SIM box fraud). They can now make long-distance calls for less money while the telcos lose out on the revenue.

Telecom arbitrage fraud: Depending on how much more expensive international calls are in one country compared to another, this fraud will result in a loss. Fraudulent businesses stand in the way of two operators. Although they route the calls through a different country with lower call rates, they pretend to be calling straight from one country.

PBX hacking: A company's personal network linking to an external phone network is called a Private Branch Exchange (PBX). This enables the organization to share lines and cut back on personnel. The PBX becomes a target for hackers who log in and use it because it is IP-based.

Access stimulation, also known as traffic pumping, occurs when local/rural exchanges increase the volume of calls to their networks in order to get the compensation charge mandated by the US FCC. Larger telecommunications are required by the Telecommunications Act of 1996 to pay a fee to the rural carriers.

Fraudsters use a stolen credit card to purchase prepaid SIM cards, smartphones, and routers from telecommunication online store. Because they are liable for issuing the chargebacks as part of the guarantee they provide, telecommunications companies lose money. This can also result in a lot of false positives.

Subscription fraud: Contract phones require a regular rent payment. The cost of the phone is covered by the monthly fee the customer agrees to pay, allowing them to utilize the new device without having to pay the whole price all at once. For high-end smartphones, scammers submit phony IDs and credit card information they have obtained through phishing, the dark web, or ID mule services. The fraudster can either pick up the phone from the store, which is simpler, or have it delivered to an address (unrelated to their genuine identity).

Smishing/SMS phishing is the practice of soliciting personal information from recipients of bulk SMS messages. Telcos don't face the majority of the costs associated with smishing, but they still don't want to be complicit in such crimes.

Wangiri fraud occurs when a fraudster leaves a missed call on your phone, leading you to return the call under the impression that it was an urgent call. The call you place typically passes through a pricey location under the fraudster's control.

Using this technique, a fraudster can take control of a customer's SIM card and use the phone number to contact the telco's customer support. They request that the service be moved to another number under their control from the customer support staff. They now have access to the customer's SMS verification information and OTPs.

Organisations enduring digital transformation must be agile to adapt to a rapidly changing business and technological environment. Now more than ever, it is essential to meet and exceed organisational expectations with a strong digital perspective supported by innovation. Future business excellence will rely heavily on granting organisations the capacity to sense, learn, react, and evolve like a living organism. This is achieved through a comprehensive but modular set of services. Live Enterprise is creating connected organisations that collaborate to innovate for the future by providing organisations with intuitive decision-making at scale, practical knowledge based on real-time options, anytime/anywhere access, and comprehensive data visibility across functions.

Material and methods

The materials used to actualize this work are as follows:

a. MATLAB Neural Toolbox
b. Java Collection Framework
c. Excel data Sheet
d. Oracle database software
e. R-packages

Methods

A high-level computational framework was utilized to design, analyse, and verify the proposed system. Fraud footprints were mapped into a new telecommunication network for end user profiling. The three approaches that were explored. They include:

1. Computational analytic modelling (CAM)

This was the work's first strategy. It provides a thorough investigation of the adaptive processes that allow for the prediction of intelligent behaviour in challenging and dynamic contexts. It centres on the computational modelling of non-linear systems such as behavioural analysis in critical domains. With CAM, the derivation of models with the capacity to learn specific task from data or experimental observation is feasible. It is a set of nature-inspired procedure to address complex real-world problems especially for complex mathematical reasoning. It normally has uncertainties during the process which could be stochastic in nature. Because fraud is a real-life problem which cannot be translated into binary language (unique values of 0 and 1) computer processing, CAM provides solutions for such problems. By leveraging probabilistic models like naïve Bayesian, linear discriminant, and neural prediction networks, this team was able to develop adaptive control actions for fraud prevention in a complex network.

2. Simulation train-classification technique (Experiment)
This is the second strategy used in this research. It uses well-crafted software libraries to simulate systemic operational behaviour in the actual world. The analytical prototype for a physical system was built and examined using simulation modelling. This makes performance prediction in actual situations easier. It examines the challenging operational conditions utilizing simulation tool libraries. Simulation modelling avoids the need to repeatedly construct numerous physical prototypes to analyse designs for new or existing parts. The derived emulated prototype is checked before making the physical version. It was possible to simulate the intricate dynamics of the fraud system model using MATLAB's neural toolbox. The physical system or procedure that will be utilized to commit the fraud is selected, its defining characteristics, mannerisms, and goal are specified, and it is then put into dynamic operation. By doing this, the difficulties of using the actual system are removed. This technique is used to demonstrate how the performance of the system is affected by the actual effects of the calculation classifier parameter/conditions. The primary simulation problems were resolved by this effort, including the collecting of reliable source data regarding the developed system selection features and behaviours.
The use of approximations and assumptions that are simplified within the simulation is another issue that is addressed. Finally, concerns about the accuracy and usefulness of the simulation's results are also addressed. Due to its lower cost and computational overhead, the stand-alone hybrid simulation technique was employed. Through the validation, the simulation fidelity (accuracy) divided into low, medium, and high categories is investigated. Although fidelity levels can be defined in a variety of ways, the following generalization was made:

Low
The very minimal amount of simulation necessary for a system to react and take inputs while producing results.

Medium
Automatically reacts to stimuli but has low precision.

High
As close as feasible to the genuine system, or virtually undetectable.

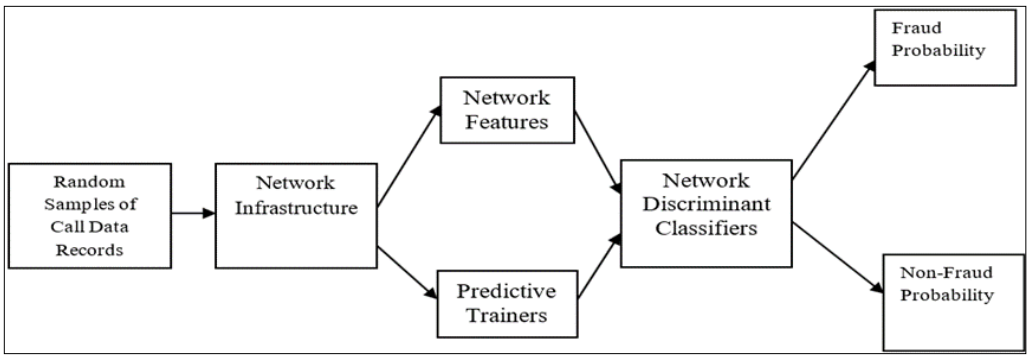

Fig1: Simulation and modelling architecture

3. Java SQL Containerization Approach (Demonstration)
It is a minimal approach to application architecture that does away with the requirement for machine virtualization. As shown in Figure 3.3, it entails placing the fraud programme inside of a Java container self-operating environment. This has benefits for putting the application into a virtual machine because the application can run without any requirements on any suitable physical system. Utilizing the free and open-source Docker Netbeans, we used containerization. The architecture of Java Docker containers runs on a variety of platforms, including bare-metal servers, virtual machines, OpenStack cloud clusters, and public instances.
The class container/data structure that contains collections of instantiated objects was examined in this work through the use of modularization. As a result, these keep things organised and in accordance with access limitations. The quantity of items (elements) a container can hold affects its size. It is possible to choose the best implementation for any given situation from among the Inherited implementations that form the foundation for a variety of container types, spanning a spectrum of sizes and complexities. The three well-established characteristics that make JCA unique are:

a. JCAaccess describes the process for gaining access to the objects within a JAVA container. The array index is used to gain access to arrays. In stacks, access follows the LIFO (last in, first out) order, while in queues, access follows the FIFO (first in, first out) order.
b. JCAstorage demonstrates a way for storing container items.
c. JCA*traversal* depicts a method For navigating through the elements of the container.

Analysis of fraud detection
This work discussed the factors of neural network simulation/emulation using a subscriber/sample radial basis neural network function (SRBNF). The computation neural controller architecture employed the clustering approach where inter-sample similarity measurement is applied to validate the subscriber traffic out of $u$samples. The designed made was to compute the hidden cluster centres, iteratively until a validation metric function, usually, mean square error (MSE) convergence is established. The composite customer attrition sources are fed into the predictive controller. Another computation engine introduced is the linear discriminant classifier. In context, the NHPP and Bayesian statistics were employed but the training algorithm used is the Naïve Bayesian network. For fraud detection, was utilized to create the neural Bayesian network for data

generation and training. The idea is the get the best mean error for accurate prediction as a result on the derived classifier index.

Data generation and collection, data pre-processing, network construction, model training, and evaluation are the primary processes utilised during design. The initial stage in developing predictive models is amassing data and cleaning samples. In this case, data were gotten from historical data of fraudulent-based and non-fraudulent subscribers. After the predictive data collection, the work used pre-processing techniques to implement the neural network training for the established number of samples (6000). The process comprises fixing the cases of an information missing, normalized data and randomized data. The missing information are replaced by the average of neighbouring values during the simulation. The work normalized the trained data before presenting the input information to the network. This is to facilitate learning accuracy by the machine learning algorithm while rejecting the variable with the smaller magnitude. The work resolved on the amount of hidden layers, neurons in each layer, transfer function in each layer, training function, weight/bias learning function, and performance function. During the training process, the weights are adjusted so as to make the actual outputs (predicated) close to the target (measured) outputs of the network. Using the classification procedure outlined previously, this work was able to ascertain the best epoch for detection using MSE accuracy for observation in the test data.

Results

The established system, comprises a Non-Homogeneous Poisson process with Bayesian neural predictive controller (Radial) linking a control call buffer in the MSC of the FD-CPMLM predictive model. The linear discriminant classifier solves the classification requirement which is deposited on the output sink. The benchmark synchronizer helps to keep a real time process. The controller has cost horizon, control horizon, weighing factor, search parameter sample time iteration. The optimal values were selected in depicting the FD-CPMLM Predictive model with Weight classifier. The neural controller used 10 hidden layers with Levenberg-Marquardt algorithm to evaluate performance of the mean square error. 200 Epochs were used, involving 14 iterations. The system generates the performance plots, training states and regression plots for the FD-CPMLM Predictive computation neural controller.

Conclusion

The results that were collected demonstrated that ENDA provided a prediction threshold for fraud attrition that was superior to DT and LR in terms of its level of reliability. Because of this, an application running ENDA is responsible for the installation of an algorithm on the servers that are part of the Fog Mobile Switching Service (FMSS) layer of the Telecommunication Service Provider (TSP) infrastructure. This algorithm ensures that an initial fraud alert is sent to users whenever a subscriber who is predicted to be fraudulent makes a call.

References

1. Al Anzi FS, Abu Zeina D. Arabic text classification using linear discriminant analysis, “International Conference on Engineering & MIS (ICEMIS), 2017, 8-10. Monastir, Tunisia. DOI: 10.1109/ICEMIS.2017.8272958.
2. Almana AM, Aksoy MS, Alzahrani RA. Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry. International Journal of Engineering Research and Applications,2014:45(1):165-171.
3. Amin A, Khan C, Ali I, Anwar S. Customer Churn Prediction in Telecommunication Industry: With and without Counter-Example. In Gelbukh A., Espinoza F.C., Galicia-Haro S.N. (eds) Nature-Inspired Computation and Machine Learning. MICAI 2014. Lecture Notes in Computer Science, 2014, 8857. Springer, Cham.
4. Amuji HO, Chukwuemeka E, Ogbuagu EM. Optimal Classifier for Fraud Detection in Telecommunication Industry. Open Journal of Optimization,2019:8:15-31. https://doi.org/10.4236/ojop.2019.81002
5. Angra S, Ahuja S. Machine learning and its applications: A review. In Proc. International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017, 23-25. Chirala, India, DOI: 10.1109/ICBDACI.2017.8070809.
6. Arif C, Kristof C, De Bock KW. A new hybrid classification algorithm for customer churns prediction based on logistic regression and decision trees. European Journal of Operational Research,2018:269(2):760-772.
7. Athmaja S, Hanumanthappa M, Vasantha K. A survey of machine learning algorithms for big data analytics”, In Proc. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, 17-18. Coimbatore, India, DOI: 10.1109/ICIIECS.2017.8276028.
8. Augustin A, Sajid A, Awais A, Muhammad N, Khalid A, Amir Hussain KH. Customer churn prediction in the telecommunication sector using a rough set approach. *Neuro computing*,2017:237(10):242-254.
9. Backiel A, Baesens B, Claeskens G. Predicting Time To-Churn of Prepaid Mobile Telephone Customers Using Social Network Analysis”, Journal of the Operational Research Society,2016:67(9):1135-1145.
10. Barlow RE, Proschan F. Statistical Theory of Reliability and Life Testing probability models. Holt, Rinehart and Winston, Inc. USA, 1975.
11. Barrack SB, Preston S. Classification and clustering for observations of event time data using non-homogeneous Poisson process models, 2018.
12. Bayindir R, Yesilbudak M, Colak M, Genc N. A Novel Application of Naive Bayes Classifier in Photovoltaic Energy Prediction. In Proc. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, 18-21. Cancun, Mexico, DOI: 10.1109/ICMLA.2017.0-108.
13. Becker RA, Volinsky C, Allan RW. Fraud Detection in Telecommunications: History and Lessons Learned. Technometrics,2009-2010:52(1):20–33. URL: http://www.tandfonline.com/doi/abs/10.1198/TECH.2009.08136, Doi:10.1198/TECH.2009.08136.
14. Berson WT, Ngai LX, Chau DCK. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications,2015:36(2):2592-2602.