Scientific Research Publishing

Open Journal of Optimization, 2019, 8, 15-31
http://www.scirp.org/journal/ojop
ISSN Online: 2325-7091
ISSN Print: 2325-7105

Optimal Classifier for Fraud Detection in Telecommunication Industry

Harrison Obiora Amuji[1*], Etus Chukwuemeka[2], Emeka Maxwel Ogbuagu[3]

[1]Department of Statistics, Federal University of Technology, Owerri, Nigeria
[2]Department of Information Management Technology, Federal University of Technology, Owerri, Nigeria
[3]Department of Math/Statistics/Computer, University of Agriculture, Makurdi, Nigeria
Email: *amujiobi@yahoo.com, chukwuemeka.etus@futo.edu.ng, emekamax84@gmail.com

How to cite this paper: Amuji, H.O., Chukwuemeka, E. and Ogbuagu, E.M. (2019) Optimal Classifier for Fraud Detection in Telecommunication Industry. *Open Journal of Optimization*, **8**, 15-31.
https://doi.org/10.4236/ojop.2019.81002

Received: September 5, 2018
Accepted: February 16, 2019
Published: February 19, 2019

Abstract

Fraud is a major challenge facing telecommunication industry. A huge amount of revenues are lost to these fraudsters who have developed different techniques and strategies to defraud the service providers. For any service provider to remain in the industry, the expected loss from the activities of these fraudsters should be highly minimized if not eliminated completely. But due to the nature of huge data and millions of subscribers involved, it becomes very difficult to detect this group of people. For this purpose, there is a need for optimal classifier and predictive probability model that can capture both the present and past history of the subscribers and classify them accordingly. In this paper, we have developed some predictive models and an optimal classifier. We 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 sample size of twenty (20) each. We obtained the prior and posterior probabilities of the groups. We group these posterior probability distributions into two sample multivariate data with two variates each. We develop linear classifier that discriminates between the genuine subscribers and fraudulent subscribers. The optimal classifier (β_{A+B}) has a posterior probability of 0.7368, and we classify the subscribers based on this optimal point. This paper focused on domestic subscribers and the parameters of interest were the number of calls per hour and the duration of the calls.

Keywords

Fraud Detection, Telecommunication, Optimal Classifier, Prior Probability, Posterior Probability

1. Introduction

Communication industry has made the world a global village, and among all

components of the industry, telecommunication is the most popular and most widely used [1]. It has created employment opportunities and empowered people economically and has removed distance, thereby saving lives and cost. Telecommunication has created opportunities for both the service providers and subscribers to do their separate but related businesses and earn their livings. But all these blessings do not come without some serious consequences of fraud in the business. Our interest in this paper is to detect fraud in the industry using the frequency and the duration of their calls. Fraud detection in telecommunication industry is vital to the survival of the industry. It is a common knowledge that fraudsters have flooded the telecommunication industries in various ways ranging from illegal access to bandwidth, attack on cyber securities, access to pocket of data, and illegal calls. All these constitute a huge loss to the telecommunication industries. These illegalities may force some of the service providers out of the industry if not properly checked. The multiplier effects of these fraudulent activities are massive loss of jobs, decline in the standard of living and its attendant consequences on those directly involved and others not directly involved. The most difficult aspect of these fraudsters is that they are smart and can hack into the data base of these service providers who should not sit back and watch them destroy their businesses. Since fraud is not localized, and does not have a permanent "office", it can be committed at anywhere and at any time. Telecommunication operators store large amounts of data related to the activities of their subscribers. In these records, there exist both normal and fraudulent records. It is expected for the fraudulent activity records to be substantially smaller than the normal activity. If it were the other way around this type of business would be impractical due to the amount of revenue lost [1].

This sector broadly has two types of users—domestic and commercial. There are cases where the connections are bought under domestic categories but the use is on a commercial scale. This causes substantial loss to the sector [2]. There is a need to adopt a data mining technique that will filter these fraudsters. Data volume has been growing at a tremendous pace due to advancements in information technology. At the same time there has been enormous development in data mining. Data mining can be defined as the process of extracting valuable information from data [3]. The telecommunication sector acquires huge amount of data due to rapidly renewable technologies, the increase in the number of subscribers and with value added services. Uncontrolled and very fast expansion of this field cause increasing losses depending on fraud and technical difficulties [4].

Today, telecommunication market all over the world is facing a severe loss of revenue due to fraudsters [5]. To overcome such business hazards and to retain the market, operators are forced to look for alternative ways of using data mining techniques and statistical tools to identify the cause in advance and to take immediate actions in response. This can be possible if the past history of the subscribers were analyzed systematically. Fortunately, telecom industries generate and maintain a large volume of data such as Call detail data and Network

data [6]. One reason for the non-utilization of this potential is the insufficient knowledge of the algorithms to be used on such data. Data mining tools and algorithms can be used to exploit the potential in the data when the data is synthesized efficiently. The advent of data mining algorithms and the development of software and hardware have led to an ease in analyzing huge and complex data [7]. Globally, the development of telecommunications industry is rapidly increasing with one innovation replacing another in a matter of years, months, and even weeks. Without doubt telecommunication is a key driver of any nation's economy. Telecommunication is the communication of information by electronic means usually over some distance. It involves the transmission and receipt of information, messages, graphics, images, voice, video and data between or among telephones, internet, satellites and radio [8].

In this area, some researchers have used different methods to determine both customer churn and fraud detection. Fraud detection and subscribers churn are related in the sense that both are concerned with subscriber's behavior. Among the models used for data mining for both churn and fraud detection are naïve Bayes model; Gaussian probability distribution; Decision Tree algorithm; logistic regression and artificial neural network (ANN). Data mining is the extraction of vital information from the bulk of data available to the telecommunication industry and using an appropriate predictive model to classify and determine the behavior of subscribers. By refining the data and building an appropriate statistical model, so much hidden information about the subscribers and service providers will be unveiled, see [9] [10] [11] [12]. This information is very vital to the survival of any service provider such as MTN, GLO, ETISALAT, MTEL, etc., in the business of telecommunication, especially in Nigeria. We shall use the subscribers' frequency of calls and the duration of such calls as parameters of interest in this paper. Then, we shall determine the prior and posterior probabilities of the subscribers and their number of calls at a given time. We shall develop a linear discriminant function which will be used to classify the posterior probability distribution into fraud and genuine subscribers. In this paper, we are concerned with statistical modeling and not machine learning or artificial intelligence method of classification.

Because of the privacy agreement between the service providers and subscribers on one hand and to protect the service providers' respective businesses on the other hand, the service providers hardly disclose their data. But nevertheless, simulation offers a close substitute for real life data. Hence, in this paper, we simulate data that depict the real life scenario and use it for the study. We simulate data on number of calls per unit time, and the call duration and our interest is on the domestic subscribers only. Eighty (80) sample data points were simulated for the study. The samples were categorized into four (4) with each having twenty (20) observations representing subscribers. The number of calls per subscriber over a period of time was also simulated and these represent real life data and are used for this study. The sample data generated from such process look like real life data drawn from a real system. We employed MINITAB 16.0 for the

simulation of the data in this work. A sample of 20 observations each on the average number of calls and rate given as follows; 8 (t = 3), 5 (t = 4), 9 (t = 12), 6 (t = 7), were simulated for the study. The values such as 8, 5,···, 7 outside the bracket represent the average number of calls per hour, and the values in bracket represent the average duration of the entire calls in minutes. Our interest is to develop a predictive data mining model for fraud detection in telecommunication industry. The simulated data were categorized into two sample multivariate data groups A and B. Most importantly, service providers determine their customers' behaviour from the nature of their current calls and their past behaviour.

2. Methodology

We need to know the history of these subscribers based on the information available to the network providers (service providers). This information is basically obtained from their call history. For this reason, the appropriate probability model that has a memory and can capture such a past history and relate it to the current history of subscribers' is the Bayesian statistic model. However Bayesian statistics requires a prior probability. Some researchers make mistake of estimating the prior probability in this type of study using a continuous distribution as though the number of calls belong to a continuous random variable. Actually, the number of calls is a Poisson problem and therefore belongs to a discrete probability distribution. The value of Poisson random variables are the non-negative integers, and any random phenomenon for which a count is of interest can be modeled by assuming a Poisson distribution, provided that the random variables satisfies certain assumptions regarding the distribution [13]. Example of such a count includes the number of telephone calls per unit time coming into the switch board of a large business. Hence, we shall estimate the prior probabilities using Poisson distribution. Since each subscriber's number of calls and time involved have non-stationery increment, we assume a non-homogenous Poisson process (NHPP) with parameter (ωt), where, ω is the call rate and t is the time duration for the calls. This has been tested and the shape parameter b was found to be greater than zero. The intensity function of power law process model ($\omega(t) = abt^{b-1}$) can be used to describe the intensity of a NHPP. The power law process model has the mean and intensity function as

$$\Lambda(t) = at^b \quad \text{and} \quad \omega(t) = \frac{\mathrm{d}\Lambda(t)}{\mathrm{d}t} = abt^{b-1} \tag{1}$$

The parameters of the model are obtained by log linear transformation of the mean value function.

$$\ln \Lambda(t) = \ln a + b \ln t \tag{2}$$

and a plot of $\ln \Lambda(t)$ against $\ln(t)$ will yield the value of $\ln(a)$ as the intercept and b as the slope of the linear graph. If the shape parameter $b = 1$, there is a stationary increment and we have HPP(ω) but for $b > 1$, we have NHPP(ω t) [14]. Hence, the predictive probability model for the priors is:

$$P_n(t) = \exp\left\{-abt^{b-1}\right\}\left[\left\{abt^{b-1}\right\}^n\right]\Big/ n!; \quad n = 0,1,\cdots \tag{3}$$

[15], where $P_n(t)$ = the probability of n number of calls at a given time (t) and the other notations retain their usual meaning as defined before.

The following assumptions must be satisfied by the random variables before we can use Equation (3) above:

Model Assumption:

The stochastic process $\{N(t), t \geq 0\}$ is called a non-homogeneous Poisson Process with rate function $\{\omega(t), t \geq 0\}$ if

1) $N(0) = 0$: (The number of events at time zero is equal to zero).

2) $\{N(t), t \geq 0\}$ has independent increment: (The number of events in non-overlapping time interval are independent).

3) $\lim_{h\to 0} \frac{o(h)}{h} = 0$: ($o(h)$—some function of smaller order than h which satisfy the limit).

4) $P\{N(t+h) = k+1 / N(t) = k\} = \lambda(t)\cdot h + 0(h)$: (The probability that exactly one event will occur in a small interval of length $t + h$ approximately equal to $\lambda(t)\cdot h + 0(h)$).

5) $P\{N(t+h) = k / N(t) = k\} = 1 - \lambda(t)\cdot h + 0(h)$: (The probability that no event occur in a small interval of length $t + h$).

6) $P\{N(t+h) = k + j / N(t) = k\} = o(h);\ j \geq 2$: (The probability that more than one event will occur in a small interval of length $t + h$).

7) The events must occur at random [16].

Bayesian statistics model is adapted for the posterior distribution since it has the attribute of capturing the prior behaviour of these subscribers to determine their current behaviour. Hence, the predictive statistical model for this study is

$$P(\psi = \zeta/\omega) = \frac{P(\zeta/\omega)P(\omega)}{\sum_{\omega=0}^{n} P(\zeta/\omega)P(\omega)} \tag{4}$$

where $P(\psi = \zeta/\omega)$ = the conditional probability that the random variable ψ assumes a specific value ζ given that its prior probability was ω. Note that ω is now a random variable. $\sum_{\omega=0}^{n} P(\zeta/\omega)P(\omega)$ = the joint probability distribution of the subscribers [17].

Our interest is to classify the subscribers as either genuine or fraudulent. Hence, this is a classification problem and linear discriminant analysis will be employed to classify the subscribers where they belong. This classification will enable service providers to determine the measures to take against these fraudsters. The discriminant analysis will discriminate between the legitimate subscribers and fraudulent ones within the network. The 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 member of similar group into the appropriate group they belong

and differentiate them from the group they do not belong to [18]. The linear discriminant function employ is given in Equation (5).

$$\beta = X^T S^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) \tag{5}$$

where $X^{\mathrm{T}} = \left(X_1 \cdots, X_p\right)$; S^{-1} is the inverse of the dispersion (variance-covariance) matrix and $\mu^{(1)} - \mu^{(2)}$ is the difference in the mean vectors between the two multivariate samples and β is the linear discriminant function. We established the optimal classifier of the discriminant function and finally classify the sample data accordingly based on their posterior probability distributions. Two multivariate sample data with two variates will be derived from Equation (4). The two sample multivariate data with two variates each are the posterior probabilities of each group. Then, we shall classify the samples as belonging to either genuine or fraudulent subscribers based on the optimal classifier (β_{A+B}). Our classification rule will be: classify the subscribers in group A into "A_1; A_2", where A_1 is the fraudulent subscribers and A_2 is the genuine subscribers. Similarly, we do the same for group B designated by "B_1; B_2". Fraud subscribers tend to make use of the services much more than the genuine subscribers and should therefore have higher probability distributions.

Definitions:

Subsc = subscribers.

n-call = the number of calls per subscriber per hour.

t (min) = the time spent on the calls.

Prop.n (n/N) = fraction of the number of calls in relation to the total number of calls.

Pr.of Prio = the probability of priors.

joint prb = the joint probabilities.

Posterior = the posterior probabilities.

Churn = the defection of subscribers from one network to another.

3. Analysis

The average number of calls and time spent in each call are presented in **Table 1**.

A plot of $\ln(t)$ against $\ln(\omega t)$ is presented in **Figure 1**.

From **Figure 1**, we found that the slope, $b = 1$. And from the relationship in Equation (2) and **Figure 1**, we have that $\ln(a) = 2.2$. Hence,

$a = \exp(2.2) = 9.0250$

The implication of the shape parameter being 1 indicate that the intensity function has stationery increment, through the PLP transformation; hence, this distribution follows HPP(ω) and the prior probability distribution of Equation (3) becomes

$$P_n(t) = \exp\{-9.0250\}\left[\{9.0250\}^n\right]/n! \tag{6}$$

Equation (6) is presented in **Table 2** by the column labeled "Prob" and Equation (4) is presented in **Table 2** by the column labeled "Poste.Pr".

Table 1. Average number of calls (ω) and average time spent (t) ($\omega = 8;\ t = 3$).

Subsc	n-call	t(min)	ωt	ln(t)	ln(ωt)
1	11	6	48	1.791759	3.871201
2	9	2	16	0.693147	2.772589
3	10	1	8	0	2.079442
4	11	1	8	0	2.079442
5	11	3	24	1.098612	3.178054
6	5	3	24	1.098612	3.178054
7	5	1	8	0	2.079442
8	7	3	24	1.098612	3.178054
9	11	2	16	0.693147	2.772589
10	6	4	32	1.386294	3.465736
11	9	2	16	0.693147	2.772589
12	8	1	8	0	2.079442
13	11	4	32	1.386294	3.465736
14	9	1	8	0	2.079442
15	8	2	16	0.693147	2.772589
16	8	2	16	0.693147	2.772589
17	10	4	32	1.386294	3.465736
18	8	1	8	0	2.079442
19	9	1	8	0	2.079442
20	7	4	32	1.386294	3.465736

Table 2. No. of calls (hr), Pn(t), Joint Prob. prior and posterior probabilities.

n-call	(a)	Exp(-a)	a^n	Prob.	Prop. N	Prior Pr.	Joint pr.	Poste. Pr
11	9.025	0.00012	3.24E+10	0.097556	0.063584	0.006203	0.112189	0.055291
9	9.025	0.00012	3.97E+08	0.131354	0.052023	0.006833	0.112189	0.06091
10	9.025	0.00012	3.58E+09	0.118547	0.057803	0.006852	0.112189	0.061079
11	9.025	0.00012	3.24E+10	0.097262	0.063584	0.006184	0.112189	0.055124
11	9.025	0.00012	3.24E+10	0.097262	0.063584	0.006184	0.112189	0.055124
5	9.025	0.00012	59873.69	0.059874	0.028902	0.00173	0.112189	0.015424
5	9.025	0.00012	59873.69	0.059874	0.028902	0.00173	0.112189	0.015424
7	9.025	0.00012	4876750	0.116113	0.040462	0.004698	0.112189	0.041878
11	9.025	0.00012	3.24E+10	0.097262	0.063584	0.006184	0.112189	0.055124
6	9.025	0.00012	540360.1	0.09006	0.034682	0.003123	0.112189	0.027841
9	9.025	0.00012	3.97E+08	0.131354	0.052023	0.006833	0.112189	0.06091
8	9.025	0.00012	44012667	0.13099	0.046243	0.006057	0.112189	0.053992
11	9.025	0.00012	3.24E+10	0.097262	0.063584	0.006184	0.112189	0.055124
9	9.025	0.00012	3.97E+08	0.131354	0.052023	0.006833	0.112189	0.06091
8	9.025	0.00012	44012667	0.13099	0.046243	0.006057	0.112189	0.053992
8	9.025	0.00012	44012667	0.13099	0.046243	0.006057	0.112189	0.053992
10	9.025	0.00012	3.58E+09	0.118547	0.057803	0.006852	0.112189	0.061079
8	9.025	0.00012	44012667	0.13099	0.046243	0.006057	0.112189	0.053992
9	9.025	0.00012	3.97E+08	0.131354	0.052023	0.006833	0.112189	0.06091
7	9.025	0.00012	4,876,750	0.116113	0.040462	0.004698	0.112189	0.041878

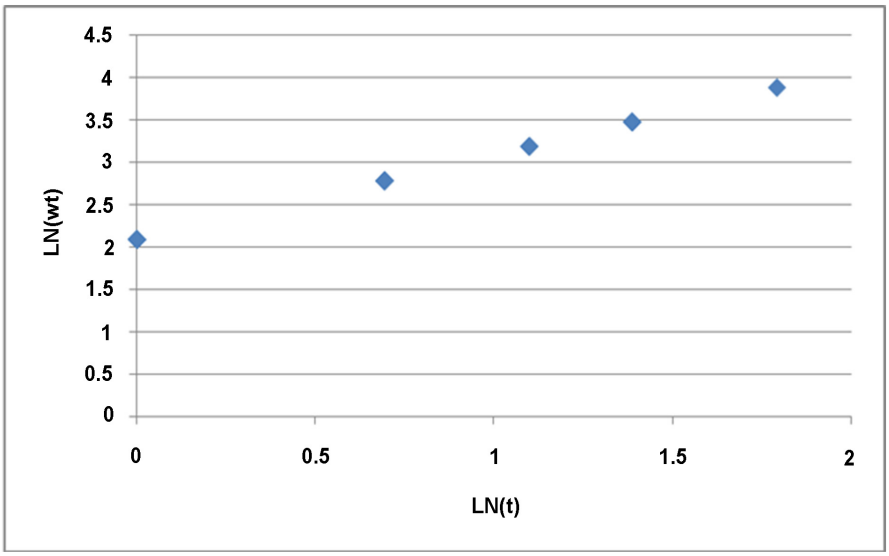

Figure 1. Graph of $\ln(t)$ against $\ln(\omega t)$.

Table 2 presents the number of calls per hour, the probability distribution, the joint probability distribution, the prior and posterior probability distributions.

The average number of calls and time spent in each call are presented in **Table 3**.

A plot of $\ln(t)$ against $\ln(\omega t)$ is presented in **Figure 2**.

From **Figure 2**, we determine the slope, $b = 1$. From the relationship in Equation (2) and **Figure 2**, we have that $\ln(a) = 1.7$. Hence, $a = \exp(1.7) = 5.4739$.

The prior probability distribution in Equation (3) becomes

$$P_n(t) = \exp\{-5.4739\}\left[\{5.4739\}^n\right]/n! \tag{7}$$

Equation (7) is presented in **Table 4** by the column labeled "Prob" and Equation (4) is presented in **Table 4** by the column labeled "Poste.Pr".

Table 4 presents the number of calls per hour, the probability distribution, the joint probability distribution, the prior and posterior probability distributions.

The average number of calls and time spent in each call are presented in **Table 5**.

A plot of $\ln(t)$ against $\ln(\omega t)$ is presented in **Figure 3**.

From **Figure 3**, we determine the slope, $b = 1$. From the relationship in Equation (2) and **Figure 3**, we have that $\ln(a) = 2.1$. Hence, $a = \exp(2.1) = 8.1662$

The prior probability distribution in Equation (3) becomes

$$P_n(t) = \exp\{-8.1662\}\left[\{8.1662\}^n\right]/n! \tag{8}$$

Equation (8) is presented in **Table 6** by the column labeled "Prob" and equation (4) is presented in **Table 6** by the column labeled "Poste.Pr".

Table 6 presents the number of calls per hour, the probability distribution, the joint probability distribution, the prior and posterior probability distributions.

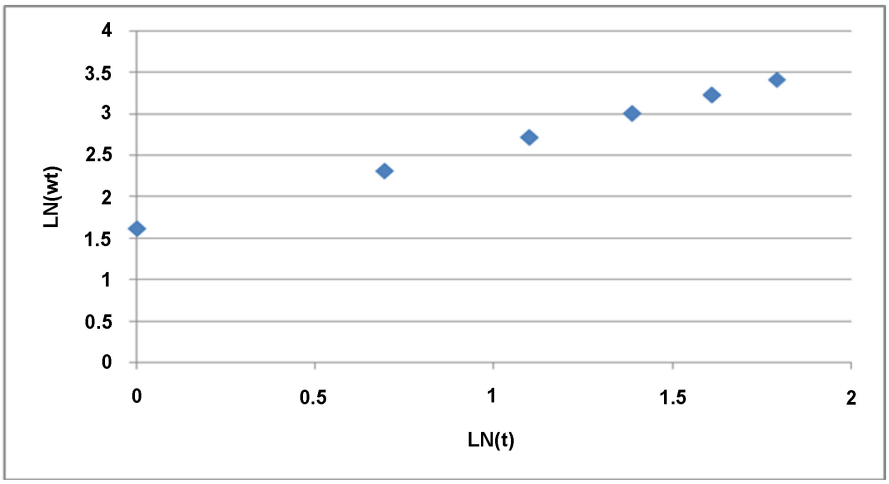

Figure 2. Graph of ln(*t*) against ln(ωt).

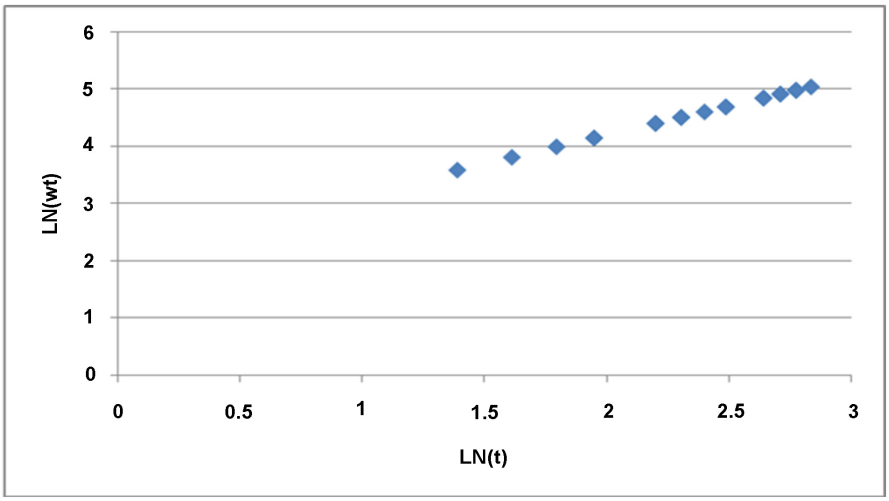

Figure 3. Graph of ln(t) against ln(ωt).

The average number of calls and time spent in each call are presented in **Table 7**.

A plot of ln(*t*) against ln(ωt) is presented in **Figure 4**.

From **Figure 4**, we determine the slope, $b = 1$. From the relationship in Equation (2) and **Figure 4**, we have that Ln(*a*) = 1.58. Hence, $a = \exp(1.58) = 4.8550$

The prior probability distribution in Equation (3) becomes

$$P_n(t) = \exp\{-4.8550\}\left[\{4.8550\}^n\right]/n! \tag{9}$$

Equation (9) is presented in **Table 8** by the column labeled "Prob" and Equation (4) is presented in **Table 8** by the column labeled "Poste.Pr".

Table 8 presents the number of calls per hour, the probability distribution, the joint probability distribution, the prior and posterior probability distributions.

Table 9 presents the posterior probability distributions for the two multivariate groups A and B.

Table 3. n-call, t(min) $\ln(t)$ and $\ln(\omega t)$ for $\omega = 5;\ t = 4$.

Subsc	n-call	t(min)	ωt	$\ln(t)$	$\ln(\omega t)$
1	8	5	25	1.609438	3.218876
2	6	5	25	1.609438	3.218876
3	9	3	15	1.098612	2.70805
4	7	1	5	0	1.609438
5	9	5	25	1.609438	3.218876
6	7	2	10	0.693147	2.302585
7	7	3	15	1.098612	2.70805
8	4	1	5	0	1.609438
9	7	3	15	1.098612	2.70805
10	5	5	25	1.609438	3.218876
11	5	6	30	1.791759	3.401197
12	7	1	5	0	1.609438
13	6	4	20	1.386294	2.995732
14	7	2	10	0.693147	2.302585
15	3	4	20	1.386294	2.995732
16	4	4	20	1.386294	2.995732
17	6	2	10	0.693147	2.302585
18	4	3	15	1.098612	2.70805
19	5	5	25	1.609438	3.218876
20	1	5	25	1.609438	3.218876

Table 4. No. of calls (hr), Pn(t), Joint Prob. prior and posterior probabilities.

n-call	(a)	Exp(−a)	a^n	Prob.	Prop. N	Prior Pr.	Joint pr.	Poste. Pr
8	5.4739	0.004195	806,073.9	0.083863	0.068376	0.005734	0.122948	0.046639
6	5.4739	0.004195	26,901.79	0.156734	0.051282	0.008038	0.122948	0.065374
9	5.4739	0.004195	4,412,368	0.051006	0.076923	0.003924	0.122948	0.031912
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
9	5.4739	0.004195	4,412,368	0.051006	0.076923	0.003924	0.122948	0.031912
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
4	5.4739	0.004195	897.8162	0.156925	0.034188	0.005365	0.122948	0.043636
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
5	5.4739	0.004195	4914.556	0.171798	0.042735	0.007342	0.122948	0.059715
5	5.4739	0.004195	4914.556	0.171798	0.042735	0.007342	0.122948	0.059715
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
6	5.4739	0.004195	26,901.79	0.156734	0.051282	0.008038	0.122948	0.065374
7	5.4739	0.004195	147,257.7	0.122564	0.059829	0.007333	0.122948	0.059642
3	5.4739	0.004195	164.0176	0.114671	0.025641	0.00294	0.122948	0.023915
4	5.4739	0.004195	897.8162	0.156925	0.034188	0.005365	0.122948	0.043636
6	5.4739	0.004195	26,901.79	0.156734	0.051282	0.008038	0.122948	0.065374
4	5.4739	0.004195	897.8162	0.156925	0.034188	0.005365	0.122948	0.043636
5	5.4739	0.004195	4914.556	0.171798	0.042735	0.007342	0.122948	0.059715
1	5.4739	0.004195	5.4739	0.022962	0.008547	0.000196	0.122948	0.001596

Table 5. n-call, t(min) $\ln(t)$ and $\ln(\omega t)$ for $\omega = 9;\ t = 12$.

Subsc	n-call	t(min)	ωt	ln(t)	$\ln(\omega t)$
1	11	14	126	2.639057	4.836282
2	11	15	135	2.70805	4.905275
3	12	10	90	2.302585	4.49981
4	13	10	90	2.302585	4.49981
5	8	7	63	1.94591	4.143135
6	12	9	81	2.197225	4.394449
7	12	15	135	2.70805	4.905275
8	10	5	45	1.609438	3.806662
9	3	11	99	2.397895	4.59512
10	4	14	126	2.639057	4.836282
11	14	17	153	2.833213	5.030438
12	12	16	144	2.772589	4.969813
13	6	4	36	1.386294	3.583519
14	14	6	54	1.791759	3.988984
15	8	12	108	2.484907	4.682131
16	9	12	108	2.484907	4.682131
17	5	16	144	2.772589	4.969813
18	12	7	63	1.94591	4.143135
19	14	9	81	2.197225	4.394449
20	8	16	144	2.772589	4.969813

Table 6. No. of calls (hr), Pn(t), Joint Prob. prior and posterior probabilities.

n-call	(a)	Exp(−a)	a^n	Prob.	Prop. N	Prior Pr.	Joint pr.	Poste. Pr
11	8.1662	0.000284	1.08E+10	0.076653	0.055556	0.004258	0.065554	0.064962
11	8.1662	0.000284	1.08E+10	0.076653	0.055556	0.004258	0.065554	0.064962
12	8.1662	0.000284	8.80E+10	0.052164	0.060606	0.003161	0.065554	0.048226
13	8.1662	0.000284	7.18E+11	0.032768	0.065657	0.002151	0.065554	0.032819
8	8.1662	0.000284	19776986	0.139349	0.040404	0.00563	0.065554	0.085887
12	8.1662	0.000284	8.80E+10	0.052164	0.060606	0.003161	0.065554	0.048226
12	8.1662	0.000284	8.80E+10	0.052164	0.060606	0.003161	0.065554	0.048226
10	8.1662	0.000284	1.32E+09	0.103253	0.050505	0.005215	0.065554	0.079549
3	8.1662	0.000284	544.5779	0.025785	0.015152	0.000391	0.065554	0.00596
4	8.1662	0.000284	4447.132	0.052642	0.020202	0.001063	0.065554	0.016223
14	8.1662	0.000284	5.87E+12	0.019113	0.070707	0.001351	0.065554	0.020616
12	8.1662	0.000284	8.80E+10	0.052164	0.060606	0.003161	0.065554	0.048226
6	8.1662	0.000284	296565.1	0.117018	0.030303	0.003546	0.065554	0.054093
14	8.1662	0.000284	5.87E+12	0.019113	0.070707	0.001351	0.065554	0.020616
8	8.1662	0.000284	19776986	0.139349	0.040404	0.00563	0.065554	0.085887
9	8.1662	0.000284	1.62E+08	0.126439	0.045455	0.005747	0.065554	0.087672
5	8.1662	0.000284	36316.17	0.085977	0.025253	0.002171	0.065554	0.03312
12	8.1662	0.000284	8.80E+10	0.052164	0.060606	0.003161	0.065554	0.048226
14	8.1662	0.000284	5.87E+12	0.019113	0.070707	0.001351	0.065554	0.020616
8	8.1662	0.000284	19776986	0.139349	0.040404	0.00563	0.065554	0.085887

Table 7. n-call, t(min) ln(t) and ln(ωt) for $\omega = 9;\ t = 12$.

Subsc	n-call	t(min)	ωt	ln(t)	ln(ωt)
1	4	2	12	0.693147	2.484907
2	4	10	60	2.302585	4.094345
3	13	7	42	1.94591	3.73767
4	8	5	30	1.609438	3.401197
5	3	7	42	1.94591	3.73767
6	5	10	60	2.302585	4.094345
7	5	4	24	1.386294	3.178054
8	5	5	30	1.609438	3.401197
9	2	1	6	0	1.791759
10	4	8	48	2.079442	3.871201
11	5	10	60	2.302585	4.094345
12	6	6	36	1.791759	3.583519
13	11	10	60	2.302585	4.094345
14	5	7	42	1.94591	3.73767
15	7	7	42	1.94591	3.73767
16	8	2	12	0.693147	2.484907
17	6	7	42	1.94591	3.73767
18	7	6	36	1.791759	3.583519
19	4	5	30	1.609438	3.401197
20	2	4	24	1.386294	3.178054

Table 8. No. of calls (hr), Pn(t), Joint Prob. prior and posterior probabilities.

n-call	(a)	Exp(-a)	a^n	Prob.	Prop. N	Prior Pr.	joint pr.	Poste. Pr
4	4.855	0.007789	555.5932	0.180321	0.035088	0.006327	0.088983	0.071104
4	4.855	0.007789	555.5932	0.180321	0.035088	0.006327	0.088983	0.071105
13	4.855	0.007789	8.33E+08	0.001042	0.035088	3.65E-05	0.088983	0.000411
8	4.855	0.007789	308683.8	0.059634	0.035088	0.002092	0.088983	0.023515
3	4.855	0.007789	114.4373	0.148565	0.035088	0.005213	0.088983	0.058583
5	4.855	0.007789	2697.405	0.175092	0.035088	0.006144	0.088983	0.069043
5	4.855	0.007789	2697.405	0.175092	0.035088	0.006144	0.088983	0.069043
5	4.855	0.007789	2697.405	0.175092	0.035088	0.006144	0.088983	0.069043
2	4.855	0.007789	23.57103	0.091801	0.035088	0.003221	0.088983	0.036199
4	4.855	0.007789	555.5932	0.180321	0.035088	0.006327	0.088983	0.071105
5	4.855	0.007789	2697.405	0.175092	0.035088	0.006144	0.088983	0.069043
6	4.855	0.007789	13095.9	0.141678	0.035088	0.004971	0.088983	0.055867
11	4.855	0.007789	35324952	0.006893	0.035088	0.000242	0.088983	0.002718
5	4.855	0.007789	2697.405	0.175092	0.035088	0.006144	0.088983	0.069043
7	4.855	0.007789	63580.6	0.098264	0.035088	0.003448	0.088983	0.038748
8	4.855	0.007789	308683.8	0.059634	0.035088	0.002092	0.088983	0.023515
6	4.855	0.007789	13095.9	0.141678	0.035088	0.004971	0.088983	0.055867
7	4.855	0.007789	63580.6	0.098264	0.035088	0.003448	0.088983	0.038748
4	4.855	0.007789	555.5932	0.180321	0.035088	0.006327	0.088983	0.071105
2	4.855	0.007789	23.57103	0.091801	0.035088	0.003221	0.088983	0.036199

Table 9. Multivariate sample data. (A) Posterior prob. from group a; (B) Posterior prob. from group B.

(A)

S/No	Variates	
1	0.055291	0.046639
2	0.06091	0.065374
3	0.061079	0.031912
4	0.055124	0.059642
5	0.055124	0.031912
6	0.015424	0.059642
7	0.015424	0.059642
8	0.041878	0.043636
9	0.055124	0.059642
10	0.027841	0.059715
11	0.06091	0.059715
12	0.053992	0.059642
13	0.055124	0.065374
14	0.06091	0.059642
15	0.053992	0.023915
16	0.053992	0.043636
17	0.061079	0.065374
18	0.053992	0.043636
19	0.06091	0.059715
20	0.041878	0.001596

(B)

S/No	Variates	
1	0.064962	0.071104
2	0.064962	0.071105
3	0.048226	0.000411
4	0.032819	0.023515
5	0.085887	0.058583
6	0.048226	0.069043
7	0.048226	0.069043
8	0.079549	0.069043
9	0.00596	0.036199
10	0.016223	0.071105
11	0.020616	0.069043
12	0.048226	0.055867
13	0.054093	0.002718
14	0.020616	0.069043
15	0.085887	0.038748
16	0.087672	0.023515
17	0.03312	0.055867
18	0.048226	0.038748
19	0.020616	0.071105
20	0.085887	0.036199

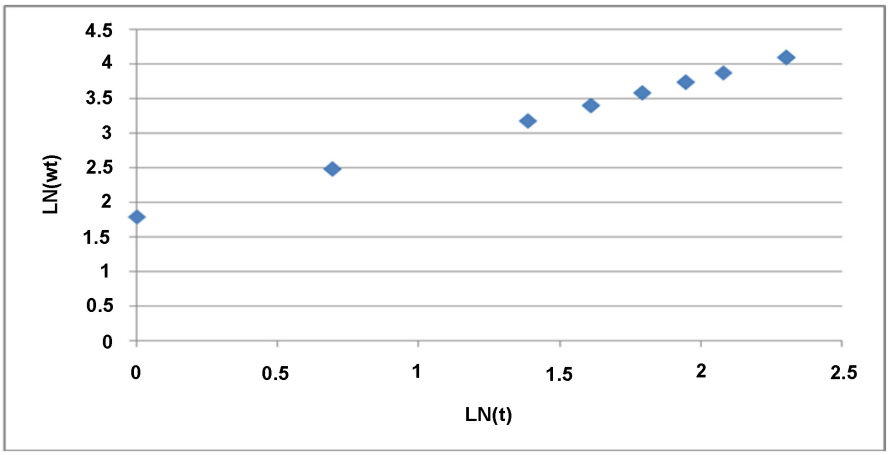

Figure 4. Graph of ln(t) against ln(ωt).

The variance-covariance matrix with two variates is given as:

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

where n_1 and n_2 respectively stand for the first and second samples respectively [16].

The variances are: $X_{11} = \sum_1^{20} X_1^2 - n\bar{X}_1^2; \quad X_{22} = \sum_1^{20} X_2^2 - n\bar{X}_2^2$

The co-variances are: $X_{12} = X_{21} = \sum_1^{20} X_1 X_2 - n\bar{X}_1\bar{X}_2;$

The observed sample multivariate data are the respective posterior probabilities of the four groups, the tendency is that their respective means will be equal thereby making the difference in the means vector to be zero. The reason behind this is that the sum of probabilities is one (1). But we can overcome this by observing the sample data carefully. Any of the sample point that cannot be approximated to two decimal places (2.d.p) with value is regarded as zero, and the sample size adjusted accordingly. Hence from sample A; delete serial number 19 from column 2; therefore, variate 1 has $n = 20$ and variate 2 has $n = 19$. Similarly, from sample B; delete serial numbers 3 and 13 from column 2; therefore, variate 1 has $n = 20$ and variate 2 has $n = 18$. The adjusted variance-covariance matrix for sample A are:

variances: $X_{11} = \sum_1^n X_1^2 - n\bar{X}_1^2;, \quad X_{22} = \sum_1^{n-1} X_2^2 - (n-1)\bar{X}_2^2$

co-variances: $X_{12} = X_{21} = \sum_1^{n-1} X_1 X_2 - \frac{n_1 + n_2}{2}\bar{X}_1\bar{X}_2;$

And

The adjusted variance-covariance matrices for sample B are:

Variances: $X_{11} = \sum_1^n X_1^2 - n\bar{X}_1^2; \quad X_{22} = \sum_1^{n-2} X_2^2 - (n-2)\bar{X}_2^2$

co-variances: $X_{12} = X_{21} = \sum_1^{n-2} X_1 X_2 - \frac{n_1 + n_2}{2}\bar{X}_1\bar{X}_2;$

The above is a symmetric matrix, where the diagonal elements are the variances. The upper and lower entries are the covariance. The values of the matrices are presented below.

From sample *A* and *B*, we have:

$$\bar{X}^{(A)} = \begin{bmatrix} 0.05 \\ 0.05255 \end{bmatrix}; \quad \bar{X}^{(B)} = \begin{bmatrix} 0.05 \\ 0.055382 \end{bmatrix}$$

$$\left(n_A - 1\right) S_A^2 = \begin{bmatrix} 0.0039 & -0.0012 \\ & 0.0029 \end{bmatrix}$$

From sample *B*, we have:

$$\left(n_B - 1\right) S_B^2 = \begin{bmatrix} 0.0129 & -0.0048 \\ & 0.0053 \end{bmatrix}$$

The pooled sample dispersion matrix is

$$\left(n_A + n_B - 2\right) S_{A+B} = \begin{bmatrix} 0.0168 & -0.00168 \\ & 0.0082 \end{bmatrix}$$

The dispersion matrix is

$$S_{A+B} = \begin{bmatrix} 0.0004541 & -0.00004541 \\ & 0.0002216 \end{bmatrix}$$

The inverse of the dispersion matrix is

$$S_{A+B}^{-1} = \begin{bmatrix} 2248.229 & 460.70423 \\ & 4607.0423 \end{bmatrix}$$

The linear discriminant function is

$$\beta = X^{\mathrm{T}} S_{A+B}^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)^{\mathrm{T}}$$

The differences in the sample mean vector for sample *A* and *B*

$$\bar{X}^{(A)} - \bar{X}^{(B)} = \begin{bmatrix} 0 \\ 0.002832 \end{bmatrix}$$

Since the variates are the posterior probabilities which cannot be negative, the difference in the mean vector cannot be negative.

$$\beta = X^{\mathrm{T}} S_{A+B}^{-1} \left(\bar{X}^{(A)} - \bar{X}^{(B)}\right)$$

$$\beta = \left(X_1 X_2\right) \begin{bmatrix} 2248.229 & 460.70423 \\ & 4607.0423 \end{bmatrix} \begin{bmatrix} 0 \\ 0.002832 \end{bmatrix}$$

$$\beta = 1.30471 x_1 + 13.0471 x_2$$

$$\beta_A = 1.30471(0.05) + 13.0471(0.05255) = 0.7509$$

$$\beta_B = 1.30471(0.05) + 13.0471(0.055382) = 0.7226$$

The optimal classifier for discrimination is

$$\beta_{A+B} = \frac{1}{2}\left(\beta_A + \beta_B\right)$$

$$\beta_{A+B} = \frac{1}{2}(0.7509 + 0.7226) = 0.7368$$

4. Classification Rule/Conclusion

Classify subscribers in group A whose posterior probability is 0.7368 and above into group A_1 and those whose posterior probability falls below 0.7368 into group A_2. Also classify subscribers in group B whose posterior probability is 0.7368 and above into group B_1 and those whose posterior probability falls below 0.7368 into group B_2.

The subscribers that belong to A_1 and B_1 are the fraudulent subscribers while those that belong to A_2 and B_2 are the legitimate subscribers. From the sample observations in **Table 9(A)** and **Table 9(B)**, all the subscribers are legitimate because their posterior probabilities are less than the optimal classifier β_{A+B}.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] João, V.C. (2014) Telecommunication Fraud Detection Using Data Mining Technique. M.Tech Thesis, Faculty of Engineering, Electrical and Computers Engineering, University of Porto, Porto.

[2] Saravanan, P., Subramaniyaswamy, V., Sivaramakrishnan, N., Arun Prakash M. and Arunkumar, T. (2014) *Contemporary Engineering Sciences*, **7**, 515-522.

[3] Amal, M.A., Mehmet, S.A. and Rasheed, A. (2014) A Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry. *International Journal of Engineering Research and Applications*, **4**, 165-171.

[4] Umman, T.Ş.G. (2010) Customer Churn Analysis in Telecommunication Sector. *Istanbul University Journal of the School of Business Administration Cilt*, **39**, 35-49.

[5] Chang, Y.-T. (2009) Applying Data Mining to Telecom Churn Management. *IJRIC*, 67-77.

[6] Balasubramanian, M. and Selvarani, M. (2014) Churn Prediction in Mobile Telecom System Using Data Mining Techniques. *International Journal of Scientific and Research Publications*, **4**, 1-5.

[7] Nabareseh, S. (2017) Predictive Analytics: A Data Mining Technique in Customer Churn Management for Decision Making. Ph.D. Dissertation, Faculty of Management and Economics, Tomas Bata University in Zlín, Zlín.

[8] Laudon, K.C., Laudon, J.P. and Brabston, M.E. (2002) Management Information System: Managing the Digital Firm. Pearson Education Canada Inc., Toronto.

[9] Alexopoulos, P. and Kafentzis, K. (2007) Towards a Generic Fraud Ontology in E Government. ICE-B, 269-276.

[10] Hollmen, J. (2000) User Profiling and Classification for Fraud Detection in Mobile Communication Networks. Ph.D. Thesis, Department of Cognitive and Computer Science and Engineering, Helsinki University of Technology, Espoo.

[11] Hiyam, A. and Tawashi, E. (2010) Detecting Fraud in Cellular Telephone Networks Jawwal Case Study. MBA Thesis, Department of Business Administration, Faculty of Commerce, Islamic University, Gaza.

[12] Kabari, L.G., Nanwin, D.N. and Nquoh, E.U. (2015) Telecommunications Subscription Fraud Detection using Artificial Neural Networks. *Transactions on Machine*

Learning and Artificial Intelligence, **3**, 19-33.

[13] Amuji, H.O., Ogbonna, C.J., Ugwuanyim, G.U., Iwu, H.C. and Nwanyibuife, O.B. (2018) Optimal Water Pipe Replacement Policy. *Open Journal of Optimization*, **7**, 41-49. https://doi.org/10.4236/ojop.2018.72002

[14] Watson, T., Colin, C., Mason, A. and Smith, M. (2001) Maintenance of Water Distribution System. 36*th Annual Conference of American Water Works Association*.

[15] Chukwu, W.I.E and Amuji, H.O. (2016) Probability, Distribution Theory and Inference. 2nd Edition, Prudent Tower Publications, Enugu, Vol. 2, 12.

[16] Chatfield, C. and Zidek, J.V. (1995) Modelling and Analysis of Stochastic Systems. Chapman and Hall, London.

[17] Arua, A.I., Chigbu, P.E., Chukwu, W.I.E., Ezekwem, C.C. and Okafor, F.C. (2000) Advanced Statistics for Higher Education. Academic Publishers, Nsukka, Vol. 1, 82.

[18] Ogbonna, C.J. and Amuji, H.O. (2018) Analysis of the Impact of Treasury Single Account on the Performance of Banks in Nigeria. *Open Journal of Statistics*, **8**, 457-467. https://doi.org/10.4236/ojs.2018.83029