International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876
Impact Factor: RJIF 5.44
www.newengineeringjournal.in
Volume 4; Issue 2; May 2018; Page No. 33-36

Review of fraud detection and churn behavior modeling techniques

Ernest O Nonum[1]*, Chukwuedozie N Ezema[2], Inyiama C Hyacinth[3]
[1] Novena University Ogume Delta State, Nigeria
[2, 3] Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

Abstract
This research pursues inductive/deductive approach by studying existing techniques for fraud detection and customer churn prediction. Telecommunication operators for instance store large amounts of data related with the activity of their clients. In these records exists both normal and fraudulent activity records. It is expected that the fraudulent activity records should 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. Furthermore, customer churn is the focal concern of most companies which are active in industries with low switching cost. Among all industries which suffer from this issue, telecommunications industry can be considered in the top of the list with approximate annual churn rate of 30%. This means wasting the money and efforts, it is like adding water to a leaking bucket". In this paper, the authors present the review of past work that has been carried out by various researchers based on fraud detection and churn behavior modeling.

Keywords: fraud detection, churn behavior modeling, customer churn

1. Introduction
Since the beginning of commercial telecommunications, the fraudsters have been causing financial damage to the companies who offered these services (Buckinx, Moons, Van Den Poel & Wets, 2016) [5]. At the start, the carriers didn't have the dimension, or the users, that they have nowadays and so the amount of fraud cases wasn't as big. This could mean that the financial damage caused, wasn't as high as it is today, but this isn't actually true. Indeed the amount of frauds was smaller but, the cost associated with a fraud attack in the beginning infrastructure was further big that it is nowadays. Later on and due to the technological advances, the cost of a fraud attack to the carrier has been decreasing, but on the opposite direction, the amount of occurrences have been increasing creating a constant financial damage (Bhattacharyya & Pendharkar, 2011) [4].
Telecommunications is wide area because it is composed by a variety of services like internet, telephones, VOIP etc. Fraud in this area is consequently an extensive subject. There are types of frauds that are characteristic of one service, like the SIM cloning fraud which is particular to the mobile phone service and there are frauds that cover a bunch of services like the subscription fraud, which is associated to the subscription of services.
Customer churn on the hand creates a huge anxiety in highly competitive service sectors especially the Telecommunications sector (Groth, 2015) [11]. The churn prediction of the mobile Telecommunication industry is on the average of 2.2% according to marketing researchers (Carrier & Powel, 2013) [7].
A number of researchers have used varied techniques to address churn in various fields. In the churn analysis and modelling of the telecommunication industry, very common algorithms deployed include Decision Trees (DT), Logistic Regression (LR), Neural Networks (NN), Naïve Bayesian Classifiers (NBC), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Self-Organising Maps (SOM), among others.

2. Review of Previous Research Work
During the last few years, there has been a lot of research in the field of fraud detection and churn prediction. Sevda and Mohammad (2015) [19] in their paper studied data mining techniques for fraud detection. They posited that scientific data mining and business intelligence technology is valuable and somewhat hidden to provide large volumes of data. The research study used service analyzer software to analyze the annual transactions related to call records of 20000 telecom subscribers. The main data mining techniques used for fraud detection (FFD) in the study was logistic models, neural networks and decision trees, all of which provide primary solutions to the problems inherent in the detection and classification of fraudulent data. The study proposed method was clustering clients based on client type. An appropriate rule for each cluster was determined the study by the behavior of group members in case of deviation from specified behavior will be known among suspected cases. The study data were based on the type of client clustering, so each cluster representing a certain type of client, the procedure will have a different behavior. The study employed decision tree algorithm and neural network model. Models are able to extract about a lot of the rules related to client behavior. Each node in the graph model was built by selecting the corresponding table; chance percent of suspected cases have been identified.
Bhattacharyya & Pendharkar (2011) [4] found in their study

that customer retention is far more economical than customer acquisition. In their work, Artificial Neural networks (ANNs) gave the best results as compared to other conventional algorithms. Furthermore, they argued that a good prediction model has to be constantly updated and should use a combination of different data mining techniques.

In another case study (Berson, Smith & Therling, 2015) [1], churn prediction and fraud detection was done using regression models, where each model comprised of different sets of variables and coefficients. A total of 6 regression models were used over a specific time period. Two models each of churner to non-churner ratio of 1:1 and 2:3 for three different analysis periods of 4, 6 and 8 months were used. The regression model with re-sampled churner to non-churner ratio of 2:3, based on data over 8 months got the best results during the testing phase. In that study, the authors concluded that due to the dynamic nature of the customer, the logistic regression models had to be updated frequently in order to achieve higher accuracy.

Bhattacharya (2011) [3] performed churn prediction and fraud detection using ANNs and decision trees. They found that the decision trees surpassed the former in terms of accuracy. They divided their project into five phases: Data Acquisition, Data preparation, Derived variables, Extracting Variables, and Model construction.

Gerpott *et al* (2001) [10] studied the German mobile market based on a sample consisting of 684 residential mobile users. They were mainly concerned with finding the relationship between the three constructs: Customer retention, loyalty and satisfaction. They showed that the three constructs are causally interlinked. In addition, they identified some other factors, which have significant effects on customer retention such as mobile service price, mobile service benefit perceptions and lack of number portability. The main limitation of this study is that it investigates customer behavioural intentions and attitudes and neglects the actual customer usage behaviours. In addition, this study employs traditional statistical techniques, which are now being questioned in the churn literature.

Kim and Yoon (2004) [14] analysed fraud detection and customer churn in the Korean mobile market based on a survey of 973 mobile users. They conducted an empirical analysis to identify the determinants of churn. Their results indicated that churn probability is associated with the level of satisfaction. Furthermore, they identified other services attributes that might affect churn including call quality, tariff level, handsets, and brand image and subscription duration. They concluded that the main cause of customer churn in the Korean mobile market is the desire to change handsets and dissatisfaction with specific service attributes, such as call quality or price level. This study shares the same limitation as Gerpott *et al*.'s (2001) [10] study, namely the use of the questionnaire as a data collection tool. The concerns about potential biases in the questionnaire-based study have been reported widely in the literature. In another study of the same market conducted by Ahn *et* (2006), the limitation of using questionnaires to collect customer data was overcome by analysing actual transactions and real customers 'data provided by a service provider. Ahn *et* (2006) investigated the key determinants of churn and reported service quality, customer usage and switching costs as the main determinants of churn.

Seo *et al* (2008) [18] analysed customer retention in the US mobile market based on a database of 31,769 customers and call log files for one of the top ten US mobile services providers. They used binary logistic regression modelling to analyse the behavioural and the demographical factors affecting customer retention. They investigated six variables: Service plan complexity, handset sophistication, length of association, connectivity quality, age and gender. As reported by Seo *et al* (2008) [18], one of the limitations of their study is that it was conducted before the introduction of local number portability. As noted earlier, the mobile market is changing rapidly and more recent and comprehensive research is needed.

Yan *et* (2012) [21] overcame the limitations of survey-based studies by using Duke Teradata's 2003 fraud detection and churn modelling tournament data, which contains 100,000 customer records extracted from a major mobile operator in the US. They developed a model called ‗Churn-Strategy Alignment Model' to evaluate churn based on 172 variables by using factor and reliability analysis. This model offers managers a new way to define customer retention strategies and helps them to understand why customers churn, rather than focusing only on who are going to churn. However, one major drawback of this model is that it fails to consider social influences.

Hwang *et* (2016) [13] investigated the Korean mobile market and suggested a lifetime value model considering propensity to churn, past financial contribution and customer potential value at the same time. They used decision tree, logistic regression and neural network to develop and evaluate their model, which can be used to segment customers and develop customer retention strategies based on the customer lifetime value. Choosing the most profitable customers and retaining them while lessening or terminating relationships with less profitable customers is one of the most successful customer retention strategies to improve business profit. One important contribution of Hwang's *et al*. (2016) [13] study is that it adds important findings to the empirical customer churn literature by considering customer lifetime value in the churn analysis. However, the study did not take into account other important factors that strongly affect customer churn, such as social influences and market characteristics.

Hung *et al* (2010) [12] used decision tree and neural networks to analyse customer churn and fraud in Taiwan based on data including customer demography data, billing, customer service interaction and call detail records data. Although the results of study show a significant improvement (based on lift chart) from those in early studies, one major criticism of this work is that it did not take account of social influences along with other factors that may affect customer churn.

Lejeune (2016) [15] applied bagging and stochastic gradient boosting (two data mining algorithms) to predict customer churn in a US mobile company. In their study, they use three groups of predictors. Behavioural, company interaction and customer demographics predictors. According to the study results, using the two algorithms performs comparably better than logistic regression. Another important finding of this study is that using ensemble classifiers produces superior

performance over single classification models. In ensemble models, multiple classification models are combined into one classifier by using different methods, such as majority voting. Nevertheless, this study shares the same limitations as others in that it does not consider social influences.

In another major study, Neslin *et al*. (2010) [17] investigated the performance of fraud detection and churn prediction models of different statistical/data mining tools. This study reported the results of a tournament in which researchers use the Duke Teradata's churn modelling data to build churn prediction models on that data. The results of this study demonstrate that logistic regression and decision tree models outperformed traditional statistical tools such as discriminant and explanatory models. The results of this study contribute to make churn prediction more accurate, but cannot provide answers for why a customer might churn (Hadden *et al* 2007).

Ling & Sheng (2016) [16] identified an important shortcoming of the existing prediction models, such as logistic regression and decision tree, which is that they lack transparency and comprehensibility. They suggested incorporating domain knowledge to improve the interpretability of the resulting models. Based on two telecom data sets they demonstrated how domain knowledge could be used to improve the interpretability of predictions models.

3. Conclusion

There is need for further research work on, "Enhanced Predictive Data Mining Algorithms for Fraud Detection and Churn Behavior Modeling in Telecommunication Systems" so as to bridges a gap in knowledge by introducing data mining as an advanced machine learning approach which is applicable in Customer relationship management realm. After the significant studies regarding the customer churn from both descriptive and predictive point of view were reviewed, the issue of data imbalance in churn datasets was discussed and remedies for it were extracted from the previous studies. As it was obvious, almost all predictive models that have been developed in this realm were utilized all or some of the RFM variables as their input variables for model building (Wei & Chiu, 2002 [20]; Coussement & Van den Poel, 2014 [9]; Hung, Yen, & Wang, 2010 [12]; Coussement & Van den Poel, 2009)[6]. There is need to conduct more study that will follow antecedents' procedure and utilize the RFM features of customer base as the input variables in clustering phase and afterwards tailor the behavioral variables proposed by Wei & Chiu (2002) [20] in order to build a robust predictive model.

We can safely conclude from the existing research in the field of fraud detection and customer churn prediction, that there is not a single model that could give the highest accuracy in all of the cases. Instead, the performance of every algorithm will differ according to the characteristics of the data. Fuerther studies should be conducted to test the conventional algorithms on data set. The study could use primary data collected from customers to create a predictive churn model that assesses customer churn rate of numerous telecommunication companies in Nigeria. This is will be unlike previous studies which were conducted in western countries. Using the IBM SPSS Modeler 18 and Rapid Miner tools, the study could generate models created by algorithms like C5.0 Decision tree algorithm, the Logistic Regression algorithm or the Discriminant Analysis algorithm. A comparative evaluation could be performed to discover the optimal model with accurate, consistent and reliable results.

Naive Bayes model classifier could be developed to do the anomaly/fraud detection on these experiments. Naive Bayes is a supervised learning classification algorithm that applies the Bayes theorem with an assumption of independency between the data features.

Many related studies focus on the step of developing prediction models and model evaluation in the data mining process, but none considered the data pre-processing step using data mining technique such as IBM SPSS Modeler 18 and Rapid Miner tools. That is, using some related technique, e.g. association rules to pre-process data may be able to improve the final performance of prediction models. As data pre-processing is an important step in the data mining process to extract useful and representative features from the original data, association rules have not been used in the data reduction step in the literature.

New research work could use three algorithms (C5.0 Decision tree algorithm, Logistic Regression algorithm and Discriminant Analysis algorithm) and performed a comparative evaluation to discover the optimal model with accurate, consistent and reliable results.

There is need for further studies to examine whether association rules can be adapted in the data pre-processing stage to reduce a large amount of information to a small and more understandable data variable in order to improve the prediction performance of using C5.0 Decision tree algorithm, Logistic Regression algorithm and Discriminant Analysis algorithm as the prediction models. The combination of the data reduction and model development steps using data mining techniques could investigated, unlike previous studies, for the problem of fraud detection and customer churn prediction.

The disadvantage of existing models and algorithms reviewed is that complex interactions among variables and attributes can affect the performance of most of the models. In addition, it becomes very complicated and difficult to visualise and interpret the models when interactions among variables and attributes become complex. Other reported disadvantages of existing models in the reviewed works are their lack of robustness and their over-sensitivity to training data sets (Burez and Van den Poel, 2009) [6].

With the ever-increasing complexity of the mobile telecommunication market, novel and efficacious algorithms such as C5.0 Decision tree, Logistic Regression and Discriminant Analysis needs to be evaluated for the purpose of customer churn prediction and fraud detection. C5.0 Decision tree algorithm, Logistic Regression algorithm and Discriminant Analysis algorithm are able to process both numerical and categorical data.

In addition, the modelling techniques and algorithms are relatively simple and have the ability to explain the relationship between input and output variables. They are easy to understand and visualize. Moreover, the decision process can be simplified to a set of business rules as demonstrated earlier (Berry & Linoff, 2016) [2]. It is a nonparametric method; therefore, no prior assumptions about the data are needed. Hence, these are potential areas for further research.

4. References

1. Berson A, Smith S, Therling K. Building data mining applications for *CRM*. New York: McGraw-Hill, 2015.
2. Berry M, Linoff G. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management (2nd Edition ed.). Indianapolis: Wiley Publishing Inc, 2016.
3. Bhattacharya C. When customers are members: customer retention in paid membership contexts. Journal of the Academy of Marketing Science. 2011; 26:31-44.
4. Bhattacharyya S, Pendharkar P. Inductive, evolutionary and neural techniques for discrimination: Acomparative study. Decision Sciences. 2011; 29:871-900.
5. Buckinx W, Moons E, Van Den Poel D, Wets G. Customer-adapted coupontargeting using feature selection. Expert Systems with Applications. 2016; 26:509-518.
6. Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. Expert System with Applications. 2009; 36:4626-4636.
7. Carrier C, Povel O. Characterizing data mining software. Intelligent Data Analysis, 2013; 7:181-192.
8. Coussement K, Van den Poel D. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, 2009; 36: 6127-6134.
9. Coussement K, Van den Poel D. integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management. 2014; 45:164-174.
10. Gerpott T, Rams W, Schindler A. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy. 2016; 25:249-269.
11. Groth R. Data Mining: Building Competitive Advantage, Santa Clara, CA: Prentice Hall, 2015.
12. Hung S, Yen D, Wang H. Applying data mining to telecom churn management. Expert Systems with Applications. 2010; 31:515-524.
13. Hwang H, Jung T, Suh E. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Exper Systems with Applications, 2016; 26:181-188.
14. Kim H, Yoon C. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy. 2016; 28:751-765.
15. Lejeune MA. Measuring the impact of data mining on churn management. Internet Research: Electronic Networking Applications and Policy. 2016; 11(5):375-387.
16. Ling C, Sheng V. Cost-sensitive learning and the class imbalance problem. Inc. Sammut, Encyclopedia of Machine Learning. Springer, 2014.
17. Neslin S, Gupta S, Kamakura W, Lu J, Mason C. Defection Detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 2010, 204-211.
18. Seo D, Ranganathan C, Babad Y. Two-level model of customer retention in the US mobile telecommunications service market. Teleccommunications Policy, 32. 182-196.
19. Sevda S, Mohammad AB. The Study of Fraud Detection in Financial and Credit Institutions with Real Data. Computer Science and Engineering. 2015; 5(2):30-36. DOI: 10.5923/j.computer.20150502.02
20. Wei C, Chiu I. Turning telecommunications call details to churn prediction: a data mining approach. Expert Systems with Applications, 2002; 23:103-112.
21. Yan L, Fassino M, Baldasare P. Predicting Customer Behavior via Calling Links. Proceedings of International Joint Conference on Neural Networks, 2012, 2555-2560. Montreal.