Journal of Engineering Science and Technology
Vol. 15, No. 4 (2020) 2289 - 2300

BROADBAND NETWORK FAULT PREDICTION USING COMPLEX EVENT PROCESSING AND PREDICTIVE ANALYTICS TECHNIQUES

EMERSON RAJA J.[1,*], HOSSEN J.,
ERVINA E. M. N.[1], TAWSIF K.[1], JESMEEN M. Z. H.[1]

[1]Faculty of Engineering and Technology, Multimedia University,
Melaka Campus, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia
*Corresponding Author: emerson.raja@mmu.edu.my

Abstract

The customer satisfaction of the broadband network mostly depends on robustness of the service offered by Internet Service Providers (ISP). Providing uninterrupted network service is essential in this communication era even though interruption in internet connection is unavoidable. However, if it is predicted earlier, the consequences can be minimized. Hence, it is essential to accurately forecast the faults in internet connection for Telecom Companies. The proposed tool for predicting broadband network fault is made up of a combination of Complex Event Processing (CEP) and Predictive Analytics (PA) techniques. The PA is used to predict network faults using techniques such as Logistic Regression (LR) or Naïve Bayes (NB). CEP is used to perform the prediction in real-time on streaming events. In this paper the performance of predictive model configured with LR is compared with the one configured with NB. Both the models had been tested for its performance using appropriate data set received from telecommunication industry using precision-recall curve and accuracy. It was found that the prediction accuracy of LR model (89.65%) is better than that of NB model (86.25%). It was also noticed that the derived AUC of LR is 0.52 which is much higher than 0.21 of NB. Hence, it was concluded that the predictive model configured with LR is performing better than the one configured with NB. So, the proposed tool configured with LR model can be implemented for fault prediction in network management systems.

Keywords: Complex event processing, Logistic regression, Machine learning, Naïve Bayes, Predictive analytics.

1. Introduction

Significant use of broadband network for communication is increasing rapidly. Information and Communication Technology (ICT) is certainly one of the important pillars for economic growth of any developing nation. The customer satisfaction of broadband network mostly depends on the robustness of the service offered by Internet Service Providers (ISP). Broadband network failure has become a critical problem for these companies [1], and to predict this failure is also challenging. It is essential to predict or forecast faults in broadband network in advance, to protect the cost, workload, customer services and reputation of the Telecommunication Companies [2].

Table 1 shows the summary of overall performance of different telecom companies (ISP) in Malaysia derived from the nationwide assessment report of year 2017 [1]. For DSL technology, service provider A failed to meet the requirement of both Round-trip time (RTT) and packet loss. Percentage of latency is calculated using the metric, RTT $\leq$ 85ms for $\geq$ 95% of the time. As for fibre, both service providers A and B were unable to meet RTT requirement with recorded measurement results of 92.23% and 91.86% respectively.

RTT is the time taken for a signal to reach the server from the end user and its acknowledgement received back from the server. Failure to attend to the requirements of more packet loss might lead to cancelling subscription plan by the customers which will negatively affect the performance ratings of the internet service provider. The main challenge in carrying out this study is dealing with the complexity of huge volume of data due to the high frequency of network sessions. The entire session data and trouble ticket data had to be pre-processed for better performance by machine learning. This was a complicated process because of the availability of vast volume of retrieved data. The internet users survey 2017 conducted by Malaysian Communications and Multimedia Commission, recorded subscriptions of 28.5 million on mobile and 2.5 million on fixed broadband categories [3].

Table 1. Overall performance of ISP during 2017 in Malaysia.

Service Provider	DSL		Fibber	
	Round Trip Time	Packet Loss	Round Trip Time	Packet Loss
A	77.34%	1.27%	92.23%	0.29%
B	99.04%	0.05%	91.86%	0.23%
C	--	--	100%	0.10%

This study enhanced the intelligence of CEP [4] by the introduction of Predictive Analytics (PA) in CEP. The PA works as an intelligent system, and the combination of CEP and PA offers better outcome for predicting events with alerts [5]. Hence, this system was employed to develop a tool for predicting broadband network fault using the combination of CEP and PA. The predictive model was configured with ML techniques such as Logistic Regression or Naïve Bayes, and the most appropriate one was then selected by applying systematic experimental tests.

The real-time prediction component is missing in some of the previous works for network fault prediction because of using batch datasets [6-8]. Correlation rules

were required to set manually before predicting the fault patterns, even though an AI model was developed using historical batch data [9]. That shows clearly that machine learning model trained with real-time data is required to avoid the necessity predefined rules imported manually. That is why in this research a proposal for network fault prediction is made using Logistic Regression method of machine learning model trained with real-time data collected from a Malaysian telecom company.

2. Methodology

Even though there are several models existing for network fault prediction, a more effective model is needed to enhance the prediction accuracy. An attempt is made in this research to improve the accuracy by using wrapper method for feature selection stage of the proposed prediction method. The three stages, pre-processing the data, designing the tool and validating the tool, in the proposed prediction method are presented in Fig. 1.

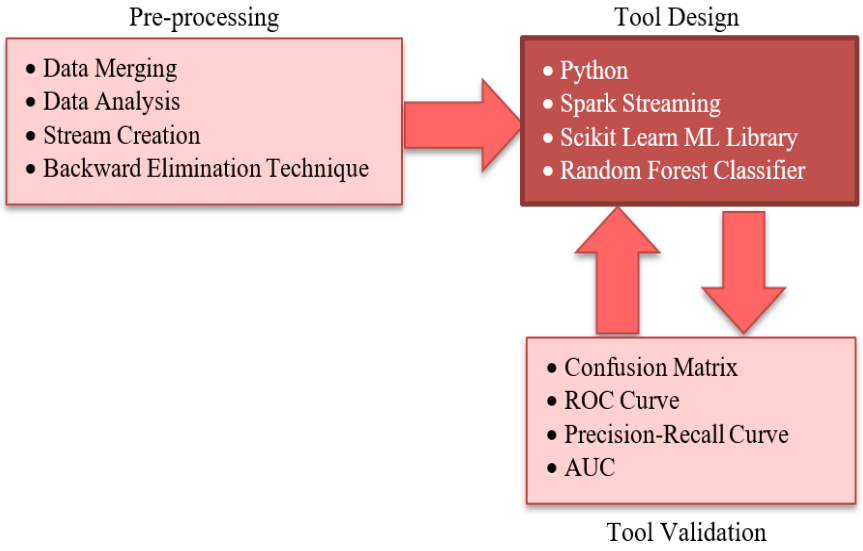

Fig. 1. Research approach.

2.1. Data collection

The quality of the signal was constantly monitored by ISPs through sending ping signal to routers/receivers at customer end. Before the customer receives a ping, it goes around different components of the network. A diagram of connected xDSL is shown in Fig. 2, where broadband remote access server (BRAS, B-RAS or BBRAS) was used to route traffic around the device on an ISP network [10]. The digital subscriber line accessed multiplexer (DSLAM) is used for gathering data traffic from different subscribers in a central socket and transferred data to a router/switch using Frame Relay, or Ethernet connection, or ATM. Finally, a connection was established with xDSL device, which totalled up Digital Subscriber Line (DSL) technology.

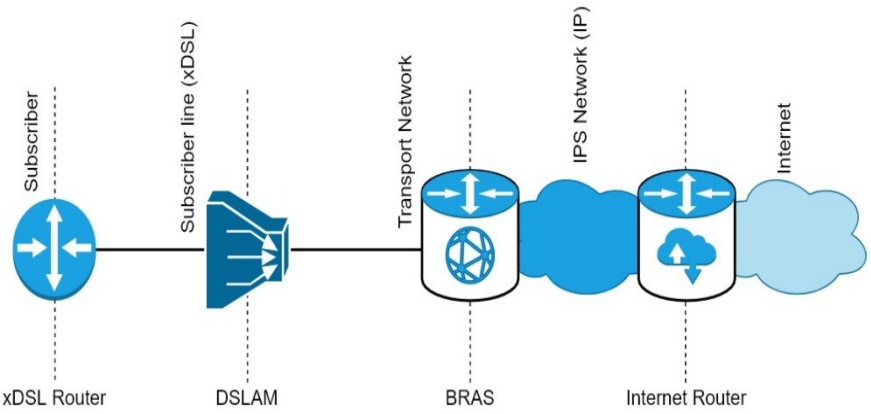

Fig. 2. xDSL connectivity diagram [11].

Information such as total uploaded bytes, total downloaded bytes, termination causes, session starting time, session ending time, total duration etc. was stored upon the termination of every session. It was estimated that a huge volume of such data is continuously stored in the database maintained by ISPs which can be used for training the network fault prediction tools. The dataset used in this study was collected from one of the ISPs from Malaysia. Connection information regarding internet sessions, user account, customer support and product details were found in this dataset. It also contained information related to events such as: user connecting to internet and terminating from internet, variations in the number of bytes during uploads or downloads and changes in the duration of sessions. The metadata was received in two files consisting of 32 columns of information about user account and RADIUS data. Incompleteness and duplication were the two major issues faced in terms of complexity in data. It was dealt with pre-processing.

2.2. Complex event processing

Complex Event Processing (CEP) technology enables extraction depending on selected set of complex events for processing a large number of data stream [12, 13]. It is useful for monitoring business activities that detect the emerging business trends of the real world. CEP has to attain high level of intelligence to enhance the intellectual performance of business processes [14]. Therefore, CEP technology was integrated with business intelligence to process large volume of data obtained from various sources in short time.

2.3. Predictive analytics

Predictive Analytics (PA) is a portion of advanced analytics technique used for predicting the unknown events which may occur in future. PA can be developed by using many different techniques (such as Datamining, Modelling, statistical, Machine Learning (ML), and Artificial Intelligence). Statistical and ML techniques were used in the development of PA for this study. Naïve Bayes (NB), and Logistic Regression (LR) algorithm were the statistical and the ML techniques used respectively in this study. NB uses "Maximum A Posteriori" decision rule for classification. The time required to train Naïve Bayes classifier is less when

compared to the time required to train other classifiers [15]. It also performs well for big datasets and high-dimensional data points. The theory of Bayesian was invented by Thomas Bayes (1702 - 1761) and named after him [16].

ML is all about determining the probabilities of an outcome as P, taking a task T and its experience E as input. LR is another technique borrowed by machine learning from the field of statistics. It is named for the function used at the core of the method, the logistic function also called the sigmoid function which can take any real-valued number and map it into a value between 0 and 1. LR is a kind of predictive analysis algorithm based on the concept of probability. The coefficients of the LR algorithm should be estimated from data used for training. This is done using maximum-likelihood estimation which makes assumptions about the distribution of training data. Since the logistic function has two different asymptotes, it can be used to divide data into "yes/no" (1 or 0) categories; the low side being "no" and the high side being "yes". LR algorithm was used to retrieve odds ratio from more than one descriptive feature. The main advantage of LR is its ability to avoid confusing properties by associating all features together.

The general training algorithm used to train LR model can be stated as follows,

Given each training data:

Step 1: Calculate a prediction using the current values of the coefficients.

Step 2: Calculate new coefficient values based on the error in the prediction.

Step 3: Repeat step 1 and 2 until the model is accurate enough (error drops to some desirable level) or for a fixed number of iterations.

2.4. System design

The CEP-PA system was designed based on the architecture shown in Fig. 3. It consisted of two sub systems, one for training and the other for prediction. Offline computations were performed by the training subsystem (upper side) using past event data gathered after being cleaned and transformed. The prediction subsystem (lower side) performed the network fault predictions based on the captured events from management system and trained the Logistic Regression model. The PA was implemented with the use of ML algorithms, for predicting fault events in a 'time-ahead-window-based-event-stream' with minimal latency.

A direct communication socket was required between the predictor and management system because of the isolation of the prediction system from management network. Within a certain amount of time after every prediction, the result was provided if the prediction made was correct or not. The result outputs and the inputs of the prediction was then used to train the model again. This is how the model kept training and continued predicting in real time.

Finally, the trained model was evaluated to prove that the trained model was effective for forecasting failures. The entire system was scripted using Python Language. Spark Streaming from Spark Framework was used for the streaming purpose. The machine learning model for the prediction was built using the Python machine learning library called Scikit Learn.

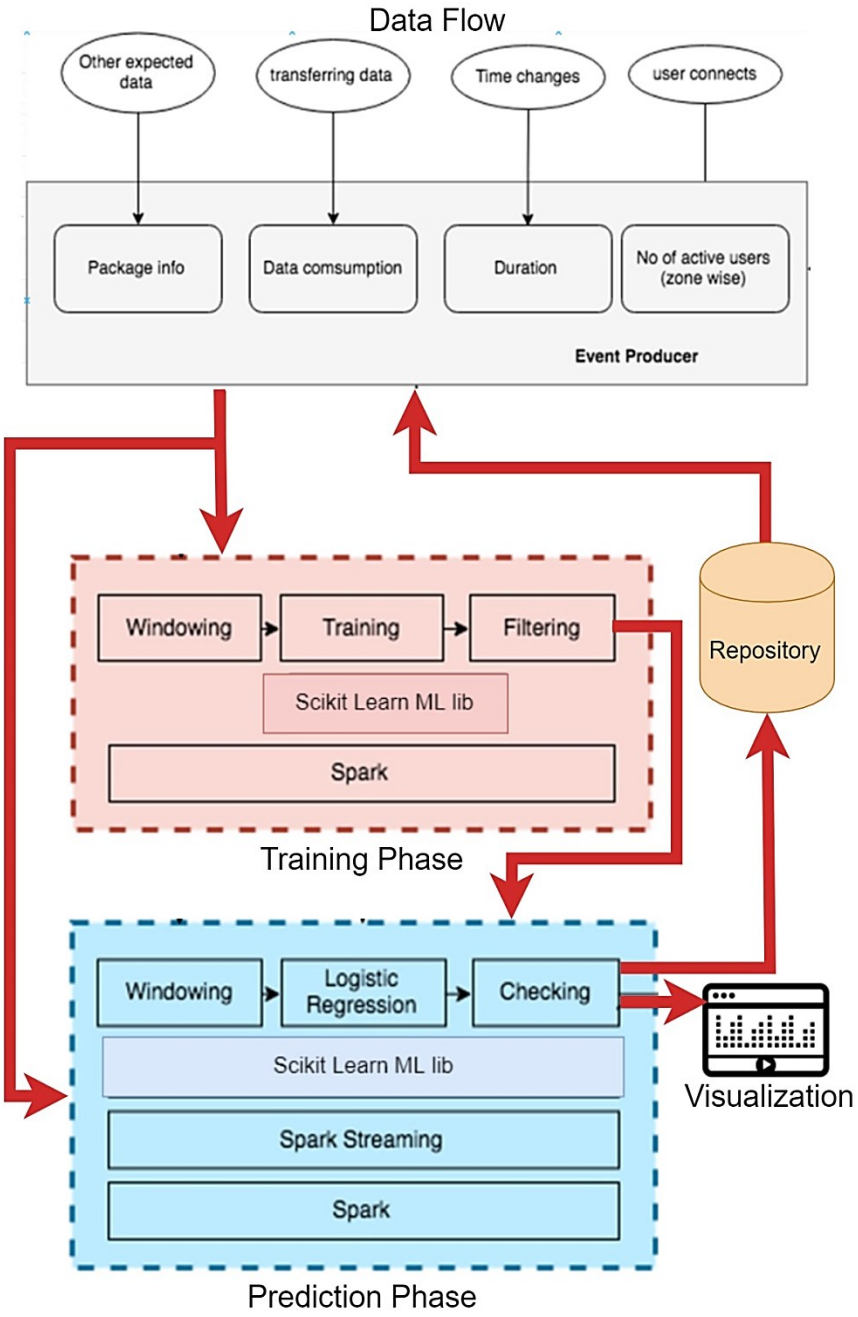

Fig. 3. CEP-PA system architecture.

2.5. Data-stream creation

It is essential to filter and clean the data for better analysis [17, 18]. Several actions were taken for cleaning the collected data such as removing the duplicate & overlapped rows and rows with NULL values in critical columns. To achieve the

aim of this study, which is to predict the network faults in real-time, the CEP technology was used. Although, CEP needs streaming data rather than batch data, the dataset collected was not streaming type. There was only session information stored at the end of every session. Since there was no streaming type of data available for predicting the future network or session faults; they had to be converted to streaming type. In order to train the ML model found in the CEP system, the entire set of data had to be split into various portions of equal size and the entire log information was also split into portions of six-hour duration.

The number of downloaded bytes and uploaded bytes got cumulated with time. However, the amount of data cumulated for every six hours was not the same. Hence, it had to be distributed equally for the purpose of training with the use of the formula shown in Eq. (1). The formula in Eq. (1) was used to calculate the number of rows (n) for a session, where duration is in seconds.

$$n = ceil\left(\frac{Duration}{6*60*60}\right) \tag{1}$$

The amount of downloaded and uploaded bytes is cumulative, and it is calculated for each row (part of a session) as F (in bytes) using the following Eq. (2).

$$F = \frac{Total\_Downloaded\_bytes}{n} \tag{2}$$

Finally, a complete set of data was generated using the accumulation process as shown in Eq. (3).

$$Fi = Fi_{-1} + F \tag{3}$$

In Eq. (3), Fi represents the amount of data to be used in ith row, each row represents a part of the entire session, which is calculated by adding the data used in previous part of the session and F. The value of i is ranging from 1 to n.

One more column, called 'active_user_in_zone' was added to the dataset as a real-time to improve prediction accuracy. The various values of this variable were mapped to zone name of the users. The value of the active user variable was increased by one whenever a user's router established a connection to internet from that zone. Similarly, the value of the active user variable was decreased whenever a user's router was disconnected from the internet. Hence, this newly inserted column reflected the active number of users at a particular zone and specific time. Therefore, this becomes `one of the most effective predictors for the proposed model, which was implemented in the form of two user defined functions. CEP engine traced the values for each zone.

2.6. Feature selection

Feature selection was needed to eliminate features with low or no predictive information. Based on different strategies and techniques, feature selection can be categorized into three distinctive groups [19]. Wrapper method was used in this study as features were evaluated using machine learning algorithm. Machine learning algorithm measured the quality of attribute and applied to each subset feature. This method helped to repeat the process until no improvement was observed and degraded the accuracy upon the removal of features [20]. Features with categorical values were transformed to numerical values using a method called String Indexer which encoded the string columns to label indices.

2.7. Evaluating techniques

The performance of the two classifiers, LR and NB, were evaluated using precision-recall curve and accuracy. Precision-recall curve was used in this research because it is a common method to compare the trained models that predict probabilities for two-class problems [15]. True Positive, False Positive, False Negative and True Negative values were used to measure accuracy, recall, precision, and F1-score of the trained models. The value for the term accuracy was a fraction which reflects the prediction accuracy of the trained model. The value for the term recall tells whether the portion of actual positives was correctly identified. The precision provided the portion of identifying positive value was actually correct. To interpret as weighted average of recall and precision values, the F1-score was used, where the value 1 is being the best and the value 0 being the worst.

3. Result and Discussion

The CEP-PA system has been trained and tested using the data stream of the selected features. The results obtained while testing the system using NB technique and LR technique for prediction are analysed and presented in this section.

3.1. Comparison of the performance of LR and NB

The true positive rate against the false positive rate is plotted in a graph, called receiver operating characteristic curve (ROC), to demonstrate the capability of the trained classifier at different threshold value. Figures 4 and 5 present the ROC curve of RF and NB respectively. The area under a ROC curve (AUC) provided accuracy measurement of the diagnostic test. The possible values of AUC range from 0.5 (no diagnostic ability) to 1.0 (perfect diagnostic ability). Since the derived AUC of LR is (0.52) much higher than that of NB (0.21), it can be safely concluded that the diagnostic ability of LR is better than NB as higher AUC values indicate better test performance.

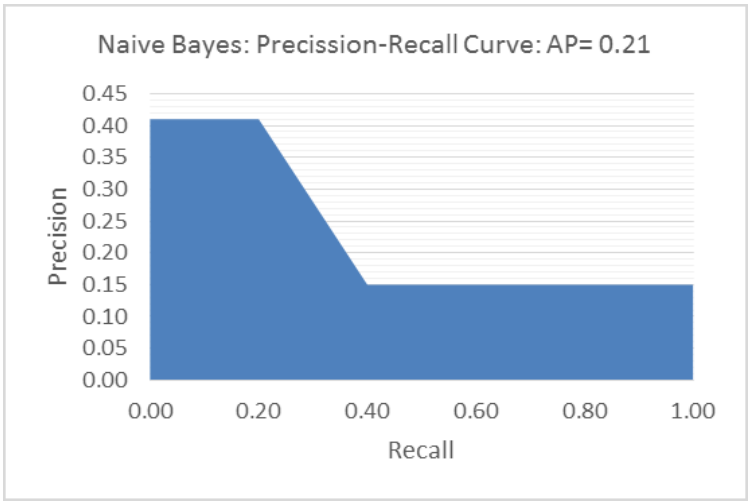

Fig. 4. Precision-recall curve of NB

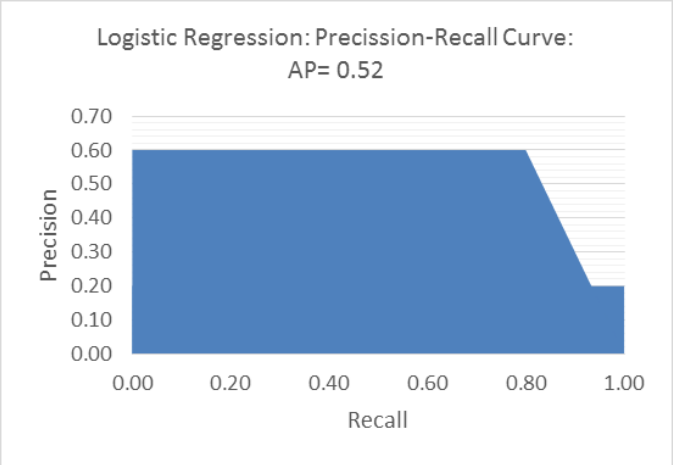

Fig. 5. Precision-recall curve of LR.

3.2. Comparing NB model and LR model for accuracy, precision, recall and F1

Confusion Matrix is presented separately for NB and LR. From the Confusion Matrixes Accuracy, Precision, Recall and F1 of NB and LR were calculated and the results are shown in Table 2. Prediction accuracy of NB and LR are also plotted in Fig. 6 for comparison. It is clear that, LR provided the highest accuracy of 89.65% in comparison to the NB algorithm with lower accuracy of 86.25%. Precision Recall and F1 of NB and LR are plotted separately in Fig. 7 for comparison. Precision Recall and F1 values of NB are 0.21, 0.24 and 0.35 respectively and LR are 0.92, 0.92 and 0.95 respectively. The graph clearly shows that LR has performed much better than NB.

Table 2. Comparison among NB and LR.

Model	Accuracy	Precision	Recall	F1
Naïve Bayes (NB)	86.25%	0.21	0.24	0.31
Logistic Regression (LR)	89.65%	0.61	0.56	0.51

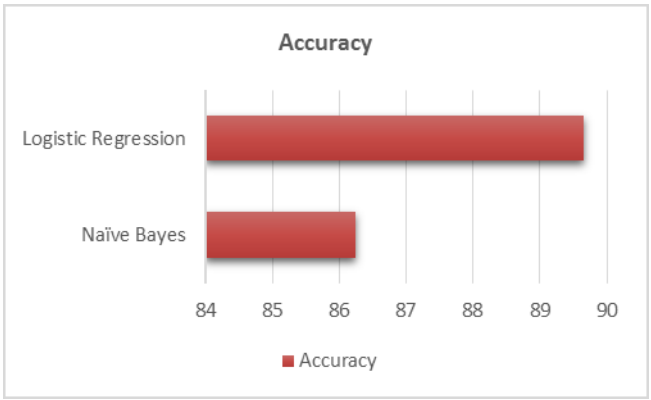

Fig. 6. Comparing accuracy of NB and LR.

Fig. 7. Comparing precesion-recall-F1 among NB and LR.

4. Conclusions

The aim of this study was to develop a suitable tool to predict the network fault. The validation of the designed tool was achieved by training and testing the selected models inside the CEP tool. The training was conducted by using the 80% of the entire data set with the selected events as features and using the remaining 20% for testing. The prediction performance of the proposed CEP tool with LR model is compared with CEP tool with NB model using accuracy, precision, recall and F1-score values. It was found that the proposed CEP tool performed better. Hence, the design with LR model instead of NB was finalized and recommended for prediction. In other words, the proposed tool configured with Logistic Regression model can be implemented for fault prediction in network management systems. Even though the proposed tool had been tested extensively and found performing well, there could be some space for further improvement. This can be achieved through training the model with additional training data set collected from different ISPs.

Acknowledgment

The research work for this paper is financially supported by the Multimedia University (MMU), Malaysia through the grant with SAP project ID MMUI/180173. We gratefully acknowledge the support of MMU without which the present study could not have been completed.

Nomenclatures

F	Amount of downloaded and uploaded bytes, bytes
F_i	Amount of data to be used in i^{th} row, bytes
n	Number of rows for a session

Abbreviations	
AUC	Area Under a ROC Curve
CEP	Complex Event Processing
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Accessed Multiplexer
ISP	Internet Service Providers
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
PA	Predictive Analytics
RADIUS	Remote Authentication Dail-In User Service
ROC	Receiver Operating Characteristic Curve
RTT	Round-Trip Time

References

1. Malaysian communication and multimedia commission. (2018). Network performance report 2017. Retrieved April 1, 2018 from https://www.skmm.gov.my/skmmgovmy/media/General/pdf/MCMC-Network-Performance-QoS-Report-2017-final-1Feb2018.pdf.
2. Jamil, J.M.; Nawawi, M.K.M.; and Ramli, R. (2016). Customer satisfaction model for mobile phone service providers in Malaysia. *Journal of Telecommunication, Electronic and Computer Engineering*, 8(8), 165-169.
3. Malaysian communications and multimedia commission. (2017). Internet users survey 2017. Retrieved April 2, 2018 from https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/MCMC-Internet-Users-Survey-2017.pdf
4. Tawsif, K.; Hossen, J.; Raja, J.E.; Jesmeen, M.Z.H.; and Arif, E.M.H. (2018). A review on complex event processing systems for big data. *Proceedings of the Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*. UPM, Malaysia, 1-6.
5. Wang, Y.; Gao, H.; and Chen, G. (2018). Predictive complex event processing based on evolving Bayesian networks. *Pattern Recognition Letters*, 105, 207-216.
6. Zhang, K.; Xu, J.; Min, M.R.; Jiang, G.; Pelechrinis, K.; and Zhang, H. (2016). Automated IT system failure prediction: a deep learning approach. *Proceedings of the IEEE International Conference on Big Data (Big Data)*. Washington, DC, 1291-1300.
7. Jaudet, M.; Iqbal, N.; and Hussain, A. (2004). Neural networks for fault-prediction in a telecommunications network. *Proceedings of the INMIC 2004: 8th International Multitopic Conference*. Lahore, Pakistan, 315-320.
8. Qian, W. (2015). Computer network fault diagnosis based on neural network. *International Journal of Future Generation Communication and Networking*, 8(5), 39-50.
9. Amador-Domínguez, E.; Serrano, E.; Manrique, D.; and Paz, J.F.D. (2019). Prediction and decision-making in intelligent environments supported by knowledge graphs, a systematic review. *Sensors*, 19(8), 1774.
10. Shrum, E. (2004). Technical report DSL forum TR-092. Retrieved January 2, 2018, from https://www.broadband-forum.org/download/TR-092.pdf.

11. Derekpsneed. (2016). What is DSLAM? Retrieved March 2, 2018, from https://www.versatek.com/blog/what-is-dslam/.
12. Luckham, D. (2002). *The power of events: an introduction to complex event processing in distributed enterprise systems*. Addison-Wesley Longman Publishing Co. Inc.
13. Eckert, M.; and Bry, F. (2010). Complex event processing (CEP). Retrieved March 3, 2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.656.2988&rep=rep1&type=pdf.
14. Hua, Z.; Gong, B.; and Xu, X. (2008). A DS-AHP approach for multi-attribute decision making problem with incomplete information. *Expert Systems with Applications*, 34(3), 2221-2227.
15. Kaur, G.; and Oberai, E.R. (2014). A review article on Naive Bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing*, 3(10), 864-868.
16. Bishop, C.M. (2006). *Pattern recognising and machine learning*. Springer-Verlag New York.
17. Jesmeen, M.Z.H.; Hossen, J.; Sayeed, S.; Rahman A.; and Arif, E.M.H. (2018). A survey on cleaning dirty data using machine learning paradigm for big data analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, 10(3), 1234-1243.
18. Hossen, J.; Jesmeen, M.Z.H.; and Sayeed, S. (2018). Modifying cleaning method in big data analytics process using random forest classifier. *Proceedings of the 7th International Conference on Computer and Communication Engineering (ICCCE)*. Kuala Lumpur, Malaysia, 208-213.
19. Fan, F.; Samworth, R.; and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10, 2013-2038.
20. Qu, G.; Hariri, S.; and Yousif, M. (2005). A new dependency and correlation analysis for features. *IEEE Transactions on Knowledge and Data Engineering*, 17(9), 1199-1207.