International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-2, December 2018

Analysis of Cloud Log Files to Prevent Account Hijacking Attacks in Hybrid Cloud

Rahul Jain, Sachin Goyal, Ratish Agrawal

Abstract: ***Cloud services supplanting the other web services in exponential rate, year-over-year. Despite the rising prevalence in cloud services, lack of proficiency in the field of cloud security is now the largest cloud challenge. The proposed method is based on the data mining techniques applied on the recorded log entries in the access log file. Before, applying data mining coarse gained log entries has been converted into fine gained log entries to improve the result. Then, generates the rule set to identify the different attacks in cloud environment. Finally, the result analysis of the proposed method has been carried out on the standard dataset through calculating the confusion matrix. Then, calculated results have been compared with other techniques through the depiction of different curves such as ROC, Lift curve, etc. These curves clear the vision about best result. Result analysis, carried out in this work shows that Logistic Regression is giving the best result among other methods.***

Keyword: Session Hijacking, Logistic Regression Decision Tree, Random Forest Roc Curve, Lift Curve

I. INTRODUCTION

The cloud computing is a transformational shift in web technology as new generation of applications. Companies adopted cloud computing for growing the business, providing the services to customers, maintaining the data in private cloud, etc. All the communication with customer or employees is conducted through creating the account to maintain the log. The accounts of the users may be hijacked by the attackers due to the vulnerabilities such as unencrypted data transmission, reuse of sequence number in virtual machine, predictable sequence number etc. Account hijacking may be exploited on two layer- application layer and TCP layer. Account session hijacking is a to seize on a session and collect the session ID of the user and dissemble as an authorized client [1]. An actual session may be created by either piracy of the token or expecting the token id. Account hijacking not only arrange the entry in the users account but it can alter integrity of user. In the today's world most trendy messaging networks are developed for performance rather than security. All the data that are exchange over the instant messaging do not encrypted all the transmission of the data go through instant messaging is not secure.

Manuscript published on 30 December 2018.
***Correspondence Author(s)
Rahul Jain, Department of Computer Science & Engineering, Radharaman Institute of Research and Technology, RGPV University, Bhopal (M.P), India
Sachin Goyal, Assistant Professor, Department of Information and Technology, University Institute of Technology, RGPV Bhopal (M.P), India
Ratish Agrawal, Assistant Professor, Department of Information and Technology, University Institute of Technology, RGPV Bhopal (M.P), India

The current information structure does not offer a method for verifying that a information truly originated from the sender. For example, a hacker may not only insert messages into an unfinished chat session, However additionally can also moreover hijack an entire session with the resource of impersonating one of the customers. It is important to research the threats of account hijacking due to the fact the outcomes can be deadly.

The rest of the paper is organized as follows: Section II discusses the related work. In Section III, proposed work has been discussed in context of the log files dataset. Section IV we discussed the results of different data mining methods and compare the results . Conclusion of the paper and future work plans is presented in Section V.

II. SURVEY ON DATA MINING TECHNIQUES

Mythili T. et al. [2] proposed a framework using logistic regression to correct prediction of heart disease. They Used Cleveland Heart Disease database. Further, they proposed a comparative study of the multiple results, which contains sensitivity, specificity, and accuracy. Jie Hao et al. [3] used the Logistic Analysis, k-Nearest Neighbor and Random Forest models to predict a past-due amount. The dataset to be they used is provided by Equifax. They calculate different parameters on the dataset and Logistic Regression gives the higher accuracy on the dataset.

Thomas et al. [4] used the logistic model approach to study the poverty of the Kenya. They survey on the income expenditure and consumption data and use Logistic regression model to estimate based on the data and determine poor and non -poor. Jai Vasanth et al. [5] used a Logistic regression model technique for find the correlation between news articles and stock prices from an information retrieval perspective. We accomplish this by ranking news articles in a very large collection based on their relevance to market price changes. The news collection is composed of six years of news and the corresponding daily stock prices. Each article in the collection is labeled as being relevant or not with respect to a significant change in stock price over a specific time. They evaluated the performance of logistic regression and find the result that lower recall our methods perform better than the baseline. S. Sivagowry et al. [6] reviews different data mining classification techniques (like Decision Tree, Naïve Bayes, Neural Network, Apriori Algorithm and MAFIA Algorithm) that are applied for diagnosing Heart Disease. It is noticed that the prediction results are powerfully stimulating and would assist physicians to do early diagnosis and make more accurate designs.

This method do not gives the hundred percentage for heart disease prediction and hence cannot be utilized solely for diagnosis. A remote health monitoring platform was designed to support heart failure severity assessment based on Classification and Regression Tree.

Artificial Neural Network algorithm is used for classifying the heart disease based on the input Learning Vector Quantization (LVQ) is a prototype based supervised classification algorithm.

Sachin Goyal et al. [7] proposed the a methodology regarding the security in the cloud, they proposed a access control system so the unauthorized user cannot access the data easily.

Sachin Goyal et al. [8] provides an advance Cross Site Scripting Attack, which could be harmful for any web site vulnerable to Cross Site Scripting Attack. After successfully implementing this attack over a vulnerable website, the attacker would be able to gain access over the content and source code of the website. The attacker will also be able to get user information from the database directory. Attackers don't have to send vulnerable link to the victim in order to steal user's data. a call for the developers, the Cross Site Scripting Attack can be more harmful as it was in the past.

III. PROPOSED WORK

We Built a preprocessing model of dataset which we download from cloud. File is in the form of log entries. We have also plan to use compression algorithm for the optimization purpose.

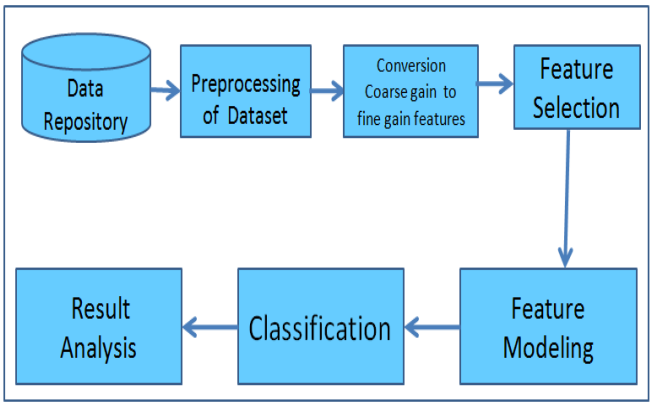

Figure 1: Block Diagram Proposed Method

Figure. 1 shown block Diagram of proposed model Our model had been passes through seven different steps. In first step we download the data from data Repository then we preprocess the dataset convert it and apply feature selection and feature modeling techniques. After these processes we apply classification model and done the result analysis of these models on different parameters.

3.1. Proposed Algorithm:

Algorithm-

Step 1: Read the one line from the access log file
Step 2: Calculate the length of current line
Step 3: If space Count is not as predefined count then got to step 1
Step 4: Else write the current line into preprocessed log file
Step 5: Extract the one log entry from the preprocessed access log file and store it into string
Step 6: Check parsed string is in the loopback address range then assign string variable user with "Admin"
Step 7: If parsed string is in range of private address range then assign string variable user with Insider.
Step 8: Else If assign string variable user with "Outsider"
Step 9: Apply Feature selection and Feature Appearance vector on preprocessed dataset.
Step 10: Apply Data mining classification techniques (Logistic Regression, Random Forest, Decision Tree) after feature selection and feature modeling on preprocessed log file dataset
Step 11: Evaluate the models performance on Confusion Matrix, ROC and Lift curve and compare the performance.

The data of log features downloaded from the cloud sources. The data contains the log features of the cloud file. We pre-process the data set and apply data mining techniques to dataset. To apply the classification models on the dataset we pre-process our target variable into two classes and apply a decision tree, random forest and logistic regression model on our dataset with R language.

3.2. Data Repository:

The standard dataset has been used for analysis, Dataset file is of approx 20 GB size that contains the access log file of cloud environment. sample dataset is represented in figure 2.

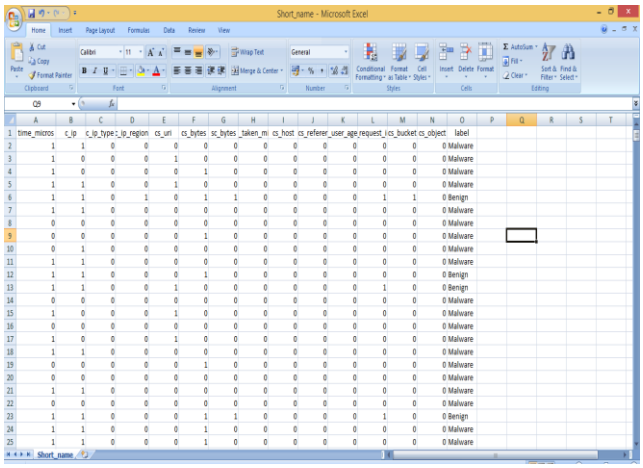

Figure 2: Classified Database

Each log entry of access file contains 'n' dimensional attribute vector (Fine grained features of log entries F: (f_1, f_2,…..,f_n), Where f_i is the value of attribute F_i. There are 2 classes: Malevolent (C_m) and Benevolent (C_b).

3.3. Pre-Processing:

The raw permissions do not arrive in a format conducive to fruitful detection. Preprocessing helps to improve the quality of data, efficiency and effectiveness. Therefore, substantial preprocessing must be applied. The preprocessing tasks applied in this work are –

- Data Cleaning and Filtering.
- Log files contains number of raw and irrelevant entries.
- Identify the duplicity of the record in access Apps.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

3.4. Feature Selection:

Feature selection is dissimilar from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset. [9]

Many databases offer applications that may download and view freely and paid. Log entries features provide the information related to the request made by the users to access the resources of end device. It may represent the behavior of resources used. To simulate the behavior of cloud has been extracted. Common features extracted from different cloud database applications.

3.5. Feature Modelling:

Selected features of the log entries have been parsed one by one and create the Feature Appearance Vector (FAV) [10]. The FAV in binary form only captures the types of feature appeared in the application or not. The ith element value in the vector is set to be 1 if the ith permission is present in the extracted permission. Otherwise, it is set to be 0 as shown in Table 1

Log Entry	Cp_ip 1 (Le_1)	Cp_ip 2 (Le_2)	Cp_ip 3 (Le_3)	time_ micros (Le 4)	Cs _byts (Le5)	Sc_byts (Le_6)	....	Cs_h ost (F_n)	Class Label
L_1	1	1	0	0	0	0	...	0	C_M
L_2	1	1	0	0	1	0	...	0	C_b
L_3	1	1	0	0	0	1	...	0	C_M
L_4	1	1	0	0	1	0	...	0	C_M
L_5	1	0	0	1	0	1	...	0	C_b
L_6	1	1	0	0	0	0	...	0	C_M
L_7	1	1	0	0	0	0	...	0	C_M

Figure 3: Database After Feature Selection

3.6. Classification Methods:

There are many classification data mining techniques but We choose logistic regression method over other methods because we have imbalanced and noisy dataset. Other data mining methods are not work well on this type of dataset. It gives very good performance on our dataset and better accuracy rate rather than models.

The constant (a_0) in the logistic regression moves the curve left and right and the slope (a_1) defines the steepness of the curve. the logistic regression equation can be written in terms of an odds ratio.

$$\frac{p}{1-p} = \exp(a_0 + ab_1x)$$

Taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors [11]. The coefficient (a1) is the amount the logit (log-odds) changes with a one unit change in x.

$$ln(\frac{p}{1-p}) = a_0 + a_1x$$

IV. RESULT ANALYSIS AND DISCUSSION

To evaluate the performance of models we used confusion Matrix , ROC curve and Lift Curves of the models . To find the best accuracy is randomly divided into 10 smaller subsets where dataset contains 612 entries, 9 subsets used for training and 1 subset is used for testing. The process is repeated 10 times for every combination. To calculate the results of different classifications models we used R language ,machine learning methods with rattle that is written in R language , the package consist inbuilt different data mining techniques we run the model and find the different accuracy parameters.

To find the results on our preprocessed log entries data set. we used three different data mining techniques random forest, logistic regression and decision tree, we run the these data mining techniques in R language with rattle data mining package and find the results, random forest gives the accuracy of 87.25% ,decision tree 87.09% and logistic regression 88.39%

Table No. 1. Comparison of Accuracies

Algorithm	Accuracy	Sensitivity	specificity
Decision Tree	87.09%	82.60%	87.45%
Random Forest	87.25%	89.74%	87.08%
Logistic Regression	88.39%	77.63%	88.16%

As we shown table no.1 that in the lift curve of three predicted model decision tree random forest and logistic regression in the chart we show glm() function has highest lift value (which defined logistic regression) in the curve when we rank the probabilities of all three model and find the curve so in our experiments LM get the highest lift rate.

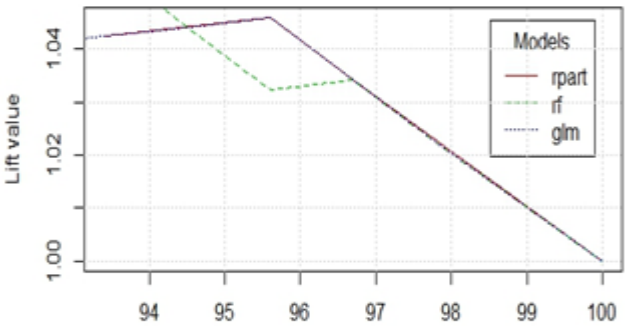

Fig .4: Comparison of Models with Lift Curve

In this work ROC curve are used because predicted probabilities of log entries are not properly classified or even when classes are highly unbalanced [12]. Roc Curve obtained by the decision tree model roc curve visualize plot between the true positive rate and the false positive rate [13].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The curve(fig no.5) shows that it has 0.62 values has under the curve. So it is a good model on our data set. In a below diagram we can see, that sensitivity at this threshold is 38% and the (1-specificity) is ~60%. To bring this curve down to a single number, we find the area under this curve (AUC)

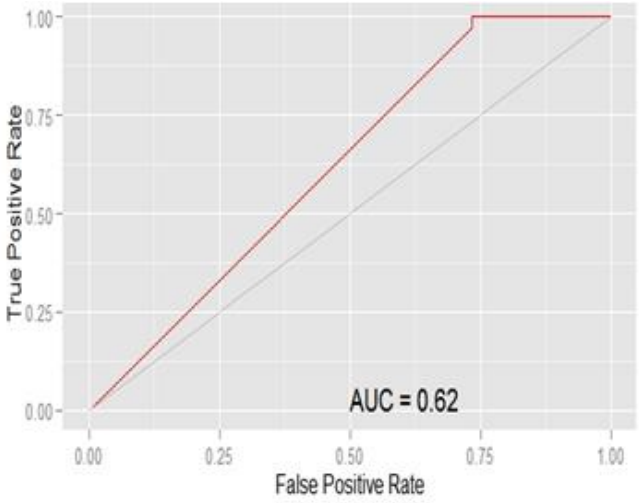

Fig .5 ROC Curve of Decision Tree Model

The curve (fig no.6) shows that it has 0.72 values has under the curve. So it is a good model on over data set. As a diagram we can see, that sensitivity at this threshold is 38.2% and the (1-specificity) is ~72%. To bring this curve down to a single number, we find the area under this curve (AUC)

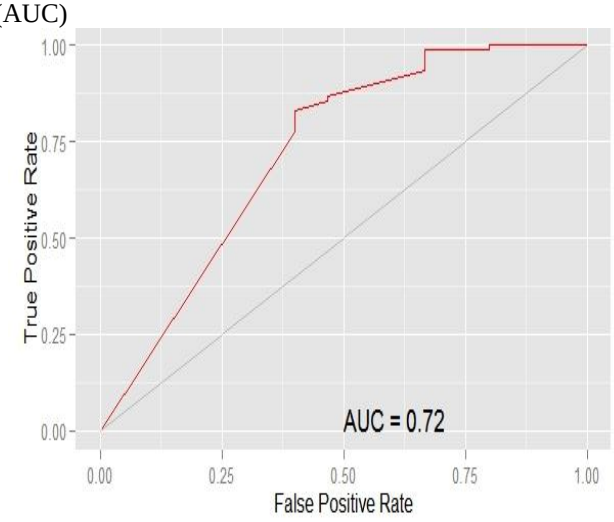

Fig. 6- ROC Curve of Random Forest Model

The curve (fig no.7) shows that it has 0.84 values has under the curve. So it is a good model on over data set. As a diagram we can see, that sensitivity at this threshold and the (1-specificity) is ~84%.To bring this curve down to a single number, we find the area under this curve (AUC)

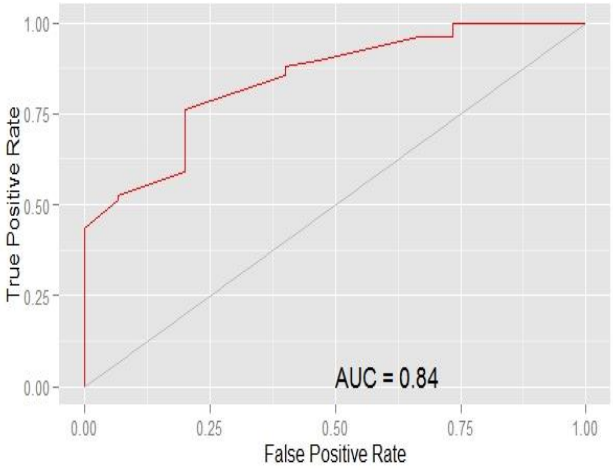

Fig 7: ROC Curve of Logistic Regression

V. CONCLUSION AND FUTURE SCOPE

To get the result we used three different data mining techniques random forest, logistic regression and decision tree, we run the these data mining techniques in R language. After applying these three data mining methods to the dataset logistic regression gives the highest accuracy of 88.39%. For the future scope we can use different non linear machine learning techniques to improve our result.

REFRENCES

1. S. Batra and A. Chhibber, "Preliminary Analysis of cloud computing & Applied Sciences," Journal of Engineering & Applied Sciences, vol. 2, no. 5, pp. 49-51, 2013.
2. Dev Mukherji, Nikita Padalia, and Abhiram Naidu Mythili T., "A Heart Disease Prediction Model using SVM-Decision," International Journal of Computer Applications (0975 – 8887), vol. Volume 68, no. No.16, April 2013.
3. Jennifer Lewis Priestley Jie Hao, "A Comparison of Machine Learning Techniques and Logistic Regression Method for the Prediction".
4. Anne Wangombe and Nancy Khadioli Thomas N O Achia, "A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data ," European Journal of Social Sciences (2010) , vol. Volume 13, no. 1, 2010.
5. Jai Vasanth Andra Ivan, "Retrieving News Articles Relevant to Stock Market Fluctuations," https://digitalcommons.kennesaw.edu/dataphdgreylit/1/.
6. M. Durairaj S. Sivagowry, International Journal of Computer Trends and Technology (IJCTT) , vol. volume 32, no. 1, February 2016.
7. Ratish Agrawal and Sachin Goyal Ankit Valdaya, "A Methology for Development and verification of Access Control system in Cloud Computing," Advanced Research in Computer and Communication Engineering, vol. 4, no. 3, March 2015.
8. Raish Agrawal Sachin goyal, "Advanced XSS (Cross Attack Scripting)Attack ," International journal of information Engineering and Electronic Busniess, vol. 7, no. 4, pp. 9-15, july 2015.
9. Girish Chandrashekarand Ferat Sahin, "A survey on feature selection methods," Computers and Electrical Engineering, pp. 16–28, 2014.
10. Muhammad Kazim and Shao Ying Zhu, "A survey on top security threats in cloud computing," International Journal of Computer Science and Applications, vol. 6, no. 3, pp. 109-113, 2015.
11. Mehrnaz Heidari Soureshjani and Ali Mohammad Kimiagari, "Calculating the best cut off point using logistic regression and neural network on credit scoring problem- A case study of a commercial bank ," African Journal of Business Management , vol. Vol. 7, no. 16, pp. pp. 1414-1421, 28 Apri 2013.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

12. Foster Provost and Jeffrey S. Simonoff Claudia Perlich, "Tree Induction vs. Logistic Regression: A Learning-Curve Analysis," Journal of Machine Learning Research , pp. 211-255, Apr. (2003).
13. A. Tekeoglu and Ali S. Tosun, "A Closer look into Privacy And Security of Chromecast Multimedia Cloud Communications," IEEE 34th Infocom-2015, pp. 121-126, 2015.

Rahul Jain, Bachelor of Engineering in Computer Science Radharaman Institute of Research and Technology RGPV Bhopal Madhya Paradesh. Presently Master of Technology in Data Science From School of Information Technology RGPV Bhopal, M.P India, Pin 462001

Sachin Goyal, Ass. Professor, Department of Information and Technology UIT RGPV Bhopal M.P, India, Pin-462001

Ratish Agrawal, Ass. Professor, Department of Information and Technology UIT , RGPV Bhopal M.P, India. Pin-462001

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)