International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-4, March 2022

Trust Prediction: Use of Decision Tree in Training & Evaluation

Archana B Saxena, Deepti Sharma, Deepshikha Aggarwal

Abstract: ***The growth of Cloud Computing is the outcome of numerous benefits associated with this technology. Like: Reduce Energy consumption, waste & carbon emission and affluence & Comfort of its working apparatus. At the same point of time its security concern is making vibrations in the research Community and regulating organizations to find the means of making it a secure and reliable stream where consumers' interests are ensured & protected. More uncertainty arises in this sector, when data theft cases are cracked up and consumers have compromised some confidential details due to intentional or unintentional action of Cloud providers. It leads the need of Trust evaluation of the provider. This trust evaluation should be in the quantitative terms and completed before selection & registration with the provider for cloud services. This paper is using predictive modeling technique to predict the trust level of the provider. Classification, a supervised Machine Learning technique that uses certification attainment status of the provider to predict its trust level.***

Keywords: Prediction, Classification, Decision Tree, Trust Prediction, Cloud, Cloud Provider, Machine Learning Algorithm.

I. INTRODUCTION

A. Related Work

Cloud Computing has become the necessity of the small and big organizations because it offers lot of benefits like: automatic updates & scaling, backup and recovery, pay as you go, no maintenance, and many more [1] [2] [3]. In the same era big renowned organizations Like Facebook, Microsoft, Marriott International, Zoom, GoDaddy, Instagram, TikTok, YouTube are dealing with the issues of data theft / breach incidents [4]. These companies have suffered losses in terms of Goodwill and monetary aspects. When big organizations can be a sufferer of data breaches incidents than others has to be little conscious while choosing cloud services. Keeping the utility of Cloud services, we cannot just rule out the use of Cloud computing but at same point of time we cannot afford to be part of Data Breach bucket. The other option is to use the secured version of the utility and be vigilant about the norms and regulations issued by the government or other regulating bodies.

Manuscript received on January 31, 2022.
Revised Manuscript received on February 05, 2022.
Manuscript published on March 30, 2022.
* Correspondence Author
Dr. Archana B Saxena*, Professor, Department of Information Technology, Jagan Institute of Management Studies, New Delhi, India.
Dr. Deepti Sharma, Associate Professor, Department of Information Technology, Jagan Institute of Management Studies, New Delhi, India.
Dr. Deepshikha Aggarwal, Professor, Department of Information Technology, Jagan Institute of Management Studies, New Delhi, India.

A new term "Trust" is also coined in reference to the cloud computing. Earlier this term is used various verticals like Marketing, Medical, Sales but now it is also denoted in concern with cloud computing between cloud consumer and cloud provider. Trust word is explored by different authors in different aspects but in authors view: Cloud consumer trust in cloud provider assures that under any normal or abnormal circumstances cloud provider will not compromise consumer's data for any kind monetary, non-monetary or personnel benefit [5]. Security and Trust are correlated in term of cloud computing. The review and analysis of concerned literature has evidenced that lot of current stream researchers have done substantial work in this aspect [6] [7] [8] [9]. Researchers have taken security as a base and try to calculate the trust value of the provider [10] [11]. Along with security various other parameters Like: SLA, Reputation, QoS, and many more aspects that are considered during trust evaluation [12] [13]. As an advancement to the same area, authors have decide to use ML technique in the Trust Prediction algorithm. Where on the basis of various certification status of the provider in various trust parameters Like Security, Governance, Audit and SLA a trust value is predicted. The dataset is generated for the providers from their public domain information. Dataset used in the current trust prediction model is based on the framework that uses the Certification attainment status of the provider to evaluate its trust value of the provider Certification Attainment - A Gizmo to Evaluate Provider's Trust: Trust Evaluation is Grounded on Provider's Attainment status Concerning Recommended Certifications.

II. METHODOLOGY

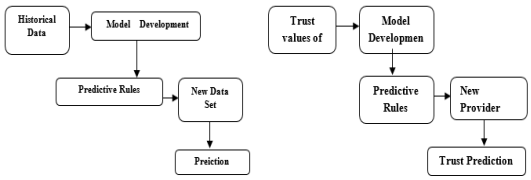

Figure 1: Predictive Modelling Process and its implementation in Trust Prediction

A. Predictive Modeling: Predictive modeling is a process that practices known outcome to craft, practice and validate a model. On the basis of this existing model future behavior or action of something or someone is predicted [14] **[Ref: Figure 1]**.

i. **Implementation in the trust prediction**: In the current piece of research authors are using OTF (Overall Trust Factor) of a provider [that is calculated on the basis of its Certification assessment status] as primary dataset [with known outcome] to practice and validate a model that will predict trust value of unknown provider.

ii. **Predictive Modeling Elements**: Predictive Modeling has two elements: Data & Modeling Technique. Modeling techniques are algorithms/process to generate the target variable on the basis of input variables. Different modeling techniques can be practiced to generate outcomes like: Statistical Mining or Machine Learning" **[Ref: figure 2]**.

Figure 2: Predictive Modeling Elements [Current Selection: Machine Learning]

Another important element of the Predictive modeling is data; the accuracy of the results is based on the consistency of the data. After data collection, before implementing it in the model data has to pass through various pre-processing stages we call it data cleaning.

B. Implementation of Elements in Trust Prediction:

a. Data: The dataset that authors have used in this prediction model is a record of 80 cloud providers. Each row of the record set represents the provider and its certification details pertaining to the trust components [7] [Ref: Table1]. The citied paper done by the authors will explain in details the reason and calculation of these values in reference to trust components like Security, Governance, SLA and Audit. It is important to mention here that certifications recommended by CSCC and CSIG in the above mentioned areas are used as a base in [15] [16] this research work. The sample table represented in the Table1 has already pass through various steps of data cleaning [17]. Although authors have done some modifications in the implementation of these data cleaning steps as per requirement of the data set collected.

Table1: Sample table used as dataset in trust prediction

Prov	Sec.	Gov.	SLA	Aud.	Trust
P1	2	1	2	1	High
P2	2	1	-	1	Med.

b. Modeling Technique: Among the two predictive modeling techniques authors have used ML (Machine Learning) in the current piece of research [Reference: Figure2]. Machine is an application of artificial Intelligence. Machine Learning is a modeling technique that can practice predictive modeling through an algorithm or automated system that improves itself automatically through experience. Among the two categories of ML algorithms: Supervised and Unsupervised, authors have chosen Supervised ML technique where Input objects and output variables both are mentioned [Reference: Figure3]. Data set used for this research work has Predictor (Input) variables like: Security, Governance, SLA and Audit and Class Label/ Output variable: Trust **[Ref: Figure4] [Ref: Table1]**.

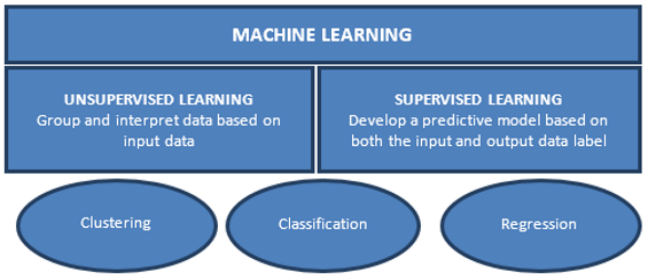

Figure3: Supervised & Unsupervised Model of Machine Learning

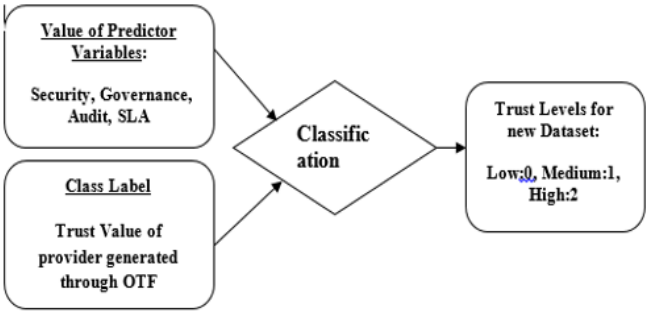

Figure4: List of Predictor variables and class Label used in Trust Prediction

c. Supervised Machine Learning Algorithm [Decision Tree]: Among the long list of supervised Machine Learning algorithms like: K-means, Random forest, Linear Regression, Logical Regression, Decision Tree and SVM (Support vector Machine). Authors have chosen "Decision Tree " to make predictions in this case as its very prevalent, very easy to understand, it generate rules, it also indicates the importance of fields and most important it accepts both numerical and categorical data. Various algorithms like C4.5, C5.0, CART, ID3 can be used to prepare Decision Tree. Authors have used ID3 algorithm for generating tree in this research.

d. Steps of constructing Decision Tree through ID3 algorithm:

i. **Entropy**: Entropy is the major of randomness in the information being processed. If the entropy is higher than it is really difficult to draw any conclusion from the information. Entropy decides how Decision tree will split the data. If the sample is homogeneous, Entropy is 0 and if sample is equally divided Entropy is 1.

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

Let there be a dataset (S) [training data] and there are C outcomes.

Ler P(I) be a proportion of S belonging to a class I, where I varies from 1 to C.

Entropy provides the information of goodness of a split. It defines the amount of information in an attribute.

Entropy(E): = $\sum_{I=1}^{C}(-p(I)log2p(I))$

Let **N** be the total no. of rows in the table

Let **E** be the entropy of the table.

Let **p,q,r** be the different feature/values of the class label.

Entropy(E)=$\sum(-p/N * log2\left(\frac{p}{N}\right) - r/N * log2\left(\frac{r}{N}\right) - q/N\ log2(q/N))$

Figure 5: Formula for calculating Entropy for the data

ii. **Information Gain:** Information gain is based on the decrease in Entropy after the data set is split on an attribute. In order to construct the "Decision Tree" we have to calculate the attribute with highest information gain. Attribute with highest information gain means it has most homogeneous branches. So the information gain is calculated for all the four attributes.

Let **N** be the total no. of rows in the table

Let **j,k,l** be the different predictor variable.

Let **m, n** be the target values of one predictor variables

Let **n** be the total occurrence of a particular feature in a predictor variable

Entropy (j)= -m/jn * log2(m/jn) – n/jn * Log2(n/jn)….

Entropy (k)= -m/kn * log2(m/kn) – n/kn * Log2(n/kn)…..

Entropy (l)= -m/ln * log2(m/ln) – n/ln * Log2(n/ln)…..

Calculating Information Gain of feature

Information Gain(Predictor_Variable): = Entropy (E) – j/N * Entropy (j) – k/N * Entropy(k) – l/N * Entropy(k)..

Figure 6: Calculation of Information for the variables of the Dataset

iii. **Tree Construction**: In order to construct the tree it is required to calculate the information gain of all four predictor variables [Ref: Figure 7]. Among all four predictor variable, Security has maximum information gain so security should be used as a root node for the tree and all three possible values of security (0,1,2) will be treated as branches of the tree [Ref figure8].

Information Gain (Security) = 0.8817516073670975 [Maximum]
Information Gain (Governance) = 0.07018755653590647
Information Gain (Audit) = 0.1678445503819601
Information Gain (SLA) = 0.0033633613131536233

Figure 7: Information gain of 4 predictor variable

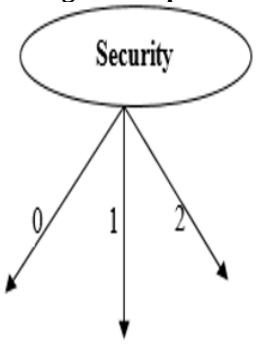

Figure 8: Level 0 Decision tree split on behalf of Root Node (Security)

iv. Which means the complete dataset is divided into three branches where security=0 or security=1 or security=2. Now after splitting we need to check the value of the target variable in concerning rows to decide for further splitting or assigning a class label. Refer Table2 that shows the records where security =2. There is only one target variable "High" means we have pure dataset and there is no need of further splitting [Ref: Figure9]. The same is checked for records where security= 1 [Ref Table3] or security=0 [Ref Table4]. The target variable is not same in both the cases so further splitting of node is required.

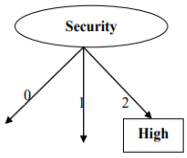

Figure 9: Decision tree level 1

Table2: Data-Frame when security=2

Index	Company	Governance	SLA	Audit	Trust-value
0	Amazon	0	0	1	2
26	Cloud stack	0	1	0	2
27	Kamatera	0	1	0	2
28	CSC	0	0	0	2

Table3: Data frame when Security=1

Index	Company	Governance	SLA	Audit	Trust-value
0	Amazon	0	0	1	2
1	IBM	1	0	0	2
2	Microsoft	0	0	1	1
3	Google	0	0	1	2

4 sample rows from a total of 34 are displayed in the output window.

Table4: Data frame when Security=0

Index	Company	Governance	SLA	Audit
7	Red Hat	0	0	0
10	SAP	1	0	1
11	Verizon	0	0	0
12	Navisite	0	0	1

4 Sample rows from a total of 38 are displayed in the output window.

So again splitting is done on the basis of remaining three attributes by calculating the information gain again. For rest three predictor variables. The same process is repeated at each level till either of the condition is true: Either the resulted records is pure dataset means all the records will have same target viable Or all the predictor variables are consumed as splitting node. The final tree for this dataset after splitting can viewed in figure 8.

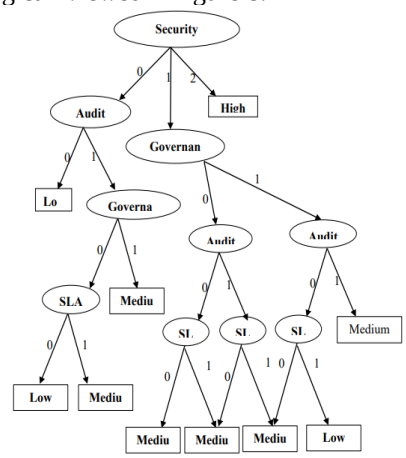

Figure 10: Level 3 Decision tree

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

III. RULES EXTRACTION:

The best part of using this decision tree is that after construction of the tree one can derive rules out of tree by looking at the nodes and branches and further these rules can be used to make predictions. On the basis of the drawn in Figure 10 one can identify various rules. Kindly refer Figure 11 for the rules that are generated from the final tree.

Rule1: If Security_Value ==2 then Trust ="HIGH"

Rule2: If Security_Value ==1 and Governance_Value==0 and Audit_Value=0 and SLA_Value ==0 then Trust= "Medium"

Rule3: If Security_Value ==1 and Governance_Value==0 and Audit_Value=0 and SLA_Value ==1 then Trust= "Medium"

Rule4: If Security_Value ==1 and Governance_Value==0 and Audit_Value=1 and SLA_Value ==0 then Trust= "Medium"

Rule5: If Security_Value ==1 and Governance_Value==0 and Audit_Value=1 and SLA_Value ==1 then Trust= "Low"

Rule6: If Security_Value ==1 and Governance_Value==1 and Audit_Value=0 and SLA_Value ==0 then Trust= "Medium"

Rule7: If Security_Value ==1 and Governance_Value==0 and Audit_Value=0 and SLA_Value ==1 then Trust= "Low"

Rule8: If Security_Value ==1 and Governance_Value==0 and Audit_Value=1 then Trust= "Medium"

Rule9: If Security_Value ==0 and Audit_Value==0 then Trust= "Low"

Rule10: If Security_Value ==0 and Audit_Value==1 and Governance==0 and SLA_Value=0 then Trust= "Low"

Rule11: If Security_Value ==0 and Audit_Value==1 and Governance==0 and SLA_Value=1 then Trust= "Medium"

Rule11: If Security_Value ==0 and Audit_Value==1 and Governance==1then Trust= "Medium"

Figure: 11 Rules generated from Decision tree

IV. RESULT ANALYSIS:

Once the tree construction is complete, it is required to check the accuracy of the algorithm and also the outputs generated by the algorithm for unknown datasets. The same has been done through Confusion Matrix. Table5 shows the confusion matrix created for the same dataset.

N= 24	Predicted LOW	Predicted MEDIUM	Predicted HIGH	The marginal sum of actuals
Actual LOW	9	0	0	9
Actual MEDIUM	3	10	0	13
Actual HIGH	0	0	2	2
The marginal sum of predictions	12	10	2	T=24

Table 5: Confusion Matrix for the unknown dataset.

With the help of confusion matrix [Ref: Table5] Accuracy Score, Misclassification Rate, Classification Report, Precision score, Recall score. All these scores indicates that the prediction algorithm generates satisfactory results [Ref: Table 6].

Table 6: Classification Report of Decision Tree Constructed On Above Mentioned Data

	Precision	Recall	F1 score	Support
0	.75	1.00	.86	9
1	1.00	.77	.87	13
2	1.00	1.00	1.00	2
Micro Average	.88	.88	.88	24
Macro Average	.92	.92	.91	24
Weighted Average	.91	.88	.88	

V. CONCLUSION AND FUTURE SCOPE

The current research piece is using Decision Tree algorithm to make predictions and results are quite satisfactory, there are many more algorithms that can be used as supervised ML algorithms like SVM (Support Vector Machine) , Random forest an many more. The authors have an intention to use the same and then compare the results generated by these different algorithms.

REFERENCES:

1. "10 Advantages of Cloud Computing for Business," [Online]. Available: https://www.salesforce.com/au/blog/2017/06/10-advantages-of-cloud-computing-for-small-businesses.html.
2. "cloud-computing-history.html," 17 April 2018. [Online]. Available: https://www.javatpoint.com/history-of-cloud-computing. [Accessed 4 May 2019].
3. "history-of-cloud-computing," 17 April 2018. [Online]. Available: https://www.computerweekly.com/feature/A-history-of-cloud-computing. [Accessed 23 May 2019].
4. "2020 Data Breaches," December 2020. [Online]. Available: https://www.identityforce.com/blog/2020-data-breaches#:~:text=May%2020%2C%202020%3A%20Over%2040,advertised%20in%20a%20hacking%20forum..
5. A. B. Saxena and M. Dawe, "Cloud Trust: A Key to attain Competitive Advantage," in ICEBM-2019," International Conference on Evidence-Based Management", BITS, Pilani, Rajasthan, 2019.
6. A. B. Saxena and M. Dawe, "Consumer's Perception on Cloud Trust: Evaluation Based on Trust Components and Sub Elements," in Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020, 2020.
7. A. B. Saxena, "Contribution of Various Components in Cloud Trust: A Cloud Consumer's Perspective," ICT for Competitive Strategies, March 2020.
8. P. Li, J. Li, Z. Huang and c. Zhi Gao, "Privacy-preserving outsourced classification in cloud computing," in Cluster computing, US, 2018.
9. S. Pearson, "Privacy, Security and Trust in Cloud Computing," in Privacy and Security for Cloud Computing, Verlag, London, 2013, pp. 3-42.
10. R. Shaikh and S. M., "Trust framework for calculating security strength of a cloud service," in 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India.
11. M. Alhanahnah, P. Bertok, Z. Tari and S. Alouneh, "Context-Aware Multifaceted Trust Framework For Evaluating Trustworthiness of Cloud Providers," Future Generation Computer Systems, pp. 488-499, 2018.
12. M. Alhamad, T. Dillon and E. Chang, "SLA-Based Trust Model for Cloud Computing," in International Conference on Network Based Information System , Takayama, 2010.
13. P. Manuel, "A Trust Model of Cloud Computing Based on Quality of Services," Annals of Operation Reserach, pp. 1-12, 2013.
14. K. Johnson and M. Kuhn, Applied Predictive Modeling, New york: Springer, 2016.
15. "Cloud Service Level Agreement Standardisation Guidelines," CSIG Memebers, Brussels, 2014.

16. A. Ali and . C. Baudoin, "Practical Guide to Cloud Computing Version 3.0," Cloud Standards Customer Council., 2017.
17. T. M. Mitchell, Machine Learning, Online: Springer, 2017.

AUTHOR PROFILE

Dr. Archana B Saxena, has attained MCA, Mphil and PHd. Degree in the field of computer science. She has vast experience of 16 years and started her professional career in teaching as Assistant Professor in ARSD (Atma Ram Sanatan Dharam) College of Delhi University. She has joined Jagan Institute of Management Studies (JIMS Rohini) in 2004. She has executed diversity of job profiles in JIMS like, Assistant Professor, Web Developer, Web administrator and currently working as Professor in JIMS. She has various publications in national and international journals Like Springer, IEEE, Elsewhere, IGI Global. she has also demonstrated her research work in various esteemed institutions like IIT Kanpur, IIT Bombay, BITS Pilani, IIM Indore, DTU(Delhi Technological University) to name a few. Other than the research topic "Trust in Cloud", her areas of interest are Research Techniques, Data Analytics, Machine Learning, and Deep Learning.

Dr. Deepti Sharma is an Associate Professor in Department of Information Technology at Jagan Institute of Management Studies, Rohini, Delhi. She has done her Phd from IGNOU in the area of "Cluster Computing". She has more than 15 years of rich teaching experience. Her research areas include Distributed Systems, Big Data Analytics, Data Sciences and Mobile Banking. She has published more than 20 research papers in various International conferences and journals.

Dr. Deepshikha Aggarwal is a highly accomplished faculty and researcher with extensive experience of over 20 years in academia. She has done B.E., M. Tech and PhD in Computer Science. She has written several research papers for various National and International journals and presented papers at different seminars and conferences. Her research interests include Social Network Analysis, Data quality, Computer networks, cyber security, E-learning and Data Science.