International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Decision Tree: A Predictive Modeling Tool Used in Cloud Trust Prediction

Archana B Saxena, Meenu Dave

Abstract: ***Trust is one of the important challenges faced by the cloud industry. Ever increasing data theft cases are contributing in worsening this issue. Regarding trust, author has a perception that this challenge can be handled to some extend if consumer can evaluate "Trust Value " of the provider or can predict the same on some reliable basis. Current research is using predictive modeling for predicting trustworthiness of cloud provider. This paper is an attempt to utilize the data mining algorithm for predictive modeling. Decision Tree, a supervised data mining algorithm has been used in the current work for making predictions. Certification attainment criteria as prime basis for trust evaluation. In current scenario, data mining algorithm will classify providers in category of low, medium and high category of trust on the basis of information displayed on the public domain.***

Keywords: Cloud, Trust, Machine Learning, Predictive Modeling, Supervised, Decision Tree.

I. INTRODUCTION & PROBLEM FORMULATION:

Cloud, an IT paradigm that have seen enormous progression since its inception. Over the years, it has become an integral part of every organization & individual's IT (Information Technology) configuration. The technology is well accepted in almost every part of globe and same trend is expected in the near future. The revenue chart of the technology confirms this notion [Reference Figure 1]. One more aspect that has gain focus along with its success is its challenges. There are many challenges [Reference: Figure 2] that technology is dealing but the three major challenges faced by the technology are: security, privacy and trust. Cloud trust is one of the imperative challenges faced by the cloud industry. Every year, data theft cases are increasing and these figures are raising concerns for its consumers [Reference Figure 3]. These data theft cases are resulting into high financial losses all over the globe. As per a survey, only India has suffered $1.77 million in 2018, which is lowest as compare to rest of the globe (Cloud Data Breach statistics n.d.). One can imagine the total financial loss whole world has to suffer because of these data theft cases. These financial aspects are motivating the researchers and forcing the legal systems and other private bodies to find a solution to this problem. Current stream researchers are trying to find some innovative solutions to solve these concerns. This paper is an attempt to solve "Trust" issue by using predictive modeling technique.

Revised Manuscript Received on August 30, 2019.
* Correspondence Author
Archana B Saxena*, Associate Professor, JIMS Rohini, Delhi, India.
Dr. Meenu Dave, Professor, Jagannath University, Jaipur, India.

Figure 1: Revenue chart of Cloud Computing
Source: (RICHMAN 2016)

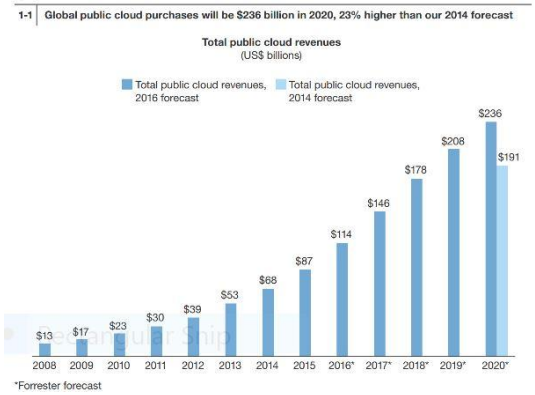


Figure 2: Cloud Computing Challenges
Source: (Zhou, et al. 2018)

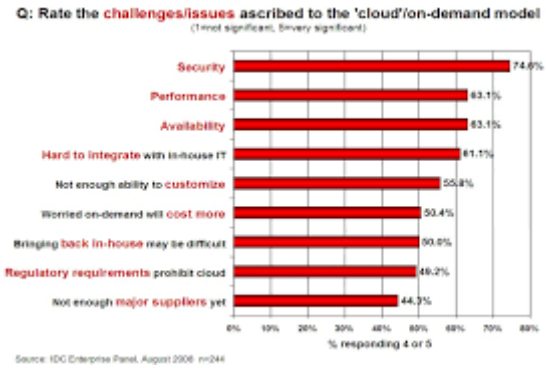
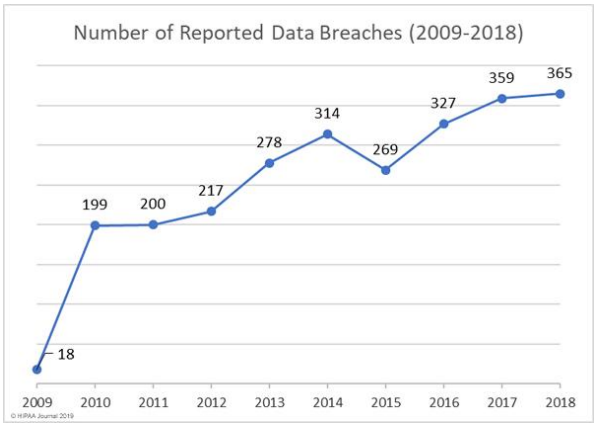


Figure 3: Reported Cloud Data breach cases of last Decade
Source:



The current work is an addition to the existing work by the author, In the order to contribute in trust issues. step in the series of n extension of a trust algorithm proposed by author [(Saxena and Dawe 2019)].

II. RESEARCH OBJECTIVE:

1. The main objective of this research piece to help the cloud consumer in predicting trustworthiness of the cloud provider on the basis of information displayed on their public domain of the provider.

III. METHODOLOGY:

Decision Tree, a supervised data mining algorithm has been used to predict the class label for the provider. The complete strategy adopted by the researcher can be explained through following steps:

1.1. **Component Selection**: In order to attain the main objective it is required to list the components that can impact Trust. There are miscellanies of elements that can impact trust. By reviewing existing concern literature, discussion with SME (Subject Matter Experts) and author's own previous knowledge, a list of components was prepared that can impact trust. Major components considered in trust evaluation are [Reference Figure 4]: Security, Governance, Audit, SLA and Diverse **(Saxena and Dawe, Loss of trust at IAAS level: Causing factors & mitigation techniques 2017).** Each component consists of some elements. The complete details of components and sub elements are explained in Table1.

Figure 4: Trust Components

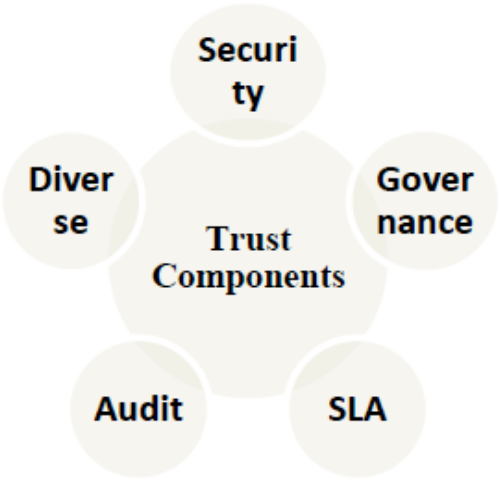


Table 1: Trust Components and sub-elements

Components	Security	Governance	SLA	Audit	Diverse
Sub Elements	Data Security	IT Governance	Performance Level Metrics	Periodical Audits	Technology Stack
	Privacy Policy	Risk Compliance	Security Metrics	Security and Management of Logs	Capacity
	Authentication & Authorization	Exit Process	Data Management Metrics	CASB Implementation	Service and Help Desk
	Cloud Network Security		Personnel Data Protection Metrics	Availability of Audit Data to Consumer	Cost and Benefit Equation
	Cloud Application Security		Responsibilities & Penalties	External Auditing	Customers Feedback
	Physical Infrastructure Security			Region and domain Concerns	

1.2. **Components Impact on Trust**: The next step is to identify the percentage impact of these components on trust. There are numerous ways we can identify the impact of these components on trust. The researcher would like to use cloud consumer's perception in this regard. In order to get cloud consumer's opinion, about the relevance of these components on trust, the researcher has used questionnaire method of survey to collect information in this regard. Online questionnaire was distributed among free and paid cloud consumers. Collected data was transferred and analyzed in SPSS version 20. **(Saxena and Dawe, IAAS Service in the Public Domain: Impact of Various Security Components on Trust 2019)**

1.3. **Percentage impact of trust components on Trust**: This information is extracted by analyzing the responses received against the online questionnaire [Reference Table 2]. On the basis of responses received against trust components, percentage contribution table was created for each component. By applying mathematical computations, weighted average table is derived for the component percentage contribution. Weighted contribution is used in OTF computation process.

Table 2: Percentage Contribution of various components in Trust

Component	% contribution in Trust	Weighted contribution
Security	57%	2
Governance	21%	1
SLA (Service Level Agreement)	10%	.5
Audit	10%	.5
Diverse	2%	.1

1.4. **Certifications and Standards recommended for these trust components**: IT industry is regulated through standards and certifications. Cloud is one of the important parts of this IT domain that is also regulated through standards and certifications. There are numerous private bodies that work for the betterment of cloud computing e.g. STAR, CSCC, CSIG. These bodies recommend various certifications to regulate various aspects of cloud. This research paper is using Standards and Certifications recommended by CSIG and CSCC for trust components and their sub elements [Ref Table 1].

1.5. **Certification attainment status of provider in relevance to these components**: In order to calculate Trust Factor for the provider, Certification and standard attainment status of the provider has been collected through public domain of the provider. A binary value is used to represent the attainment status:

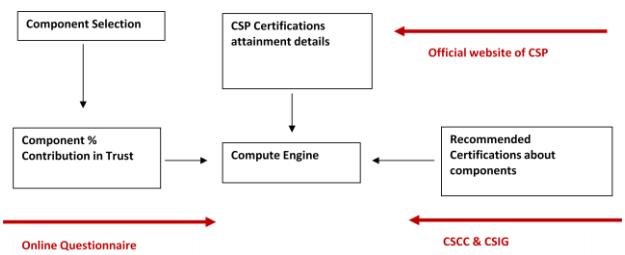
0: absence of recommended Certification or standard

1: Presence of recommended Certification and standard

1.6. **Trust Evaluation Model**: Author has proposed a trust model [Reference Figure 1] that can compute trust value of provider on the basis of the above mentioned components.

The availability status of the provider, regarding recommended standards and certifications was the prime criteria of trust evaluation in this model. The recommendations made by CSCC and CSIG will be considered in this model **(Saxena and Dawe, Certification Attainment - A Gizmo to Evaluate Provider's Trust 2019)**.

Figure 5: Execution of Trust Model



1.7. **Trust Computation (OTF)**: The results obtained in points a, b and c was inserted into the model and Overall Trust Value (OTF) of the provider is generated by using formulas from Equation 1 to Equation 5. Diverse, one of the trust components was not used in trust computation as there are no recommended Certifications and standards available for this component and its percentage contribution in trust is also very low.

Security Value (Sv) $SV = \sum_{i=1}^{6} \int CSi * CWi$

Equation 1

Governance Value (Gv)= $GV = \sum_{j=1}^{3} \int CGj * CWj$

Equation 2

SLA Value (SLAv) = $SL = \sum_{k=1}^{4} \int CSLk * CWk$

Equation 3

Audit Value (Av)= $A = \sum_{l=1}^{5} \int CAl * CWl$.

Equation 4

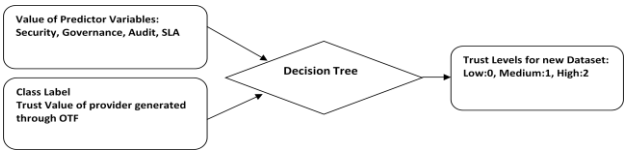
OTF (Overall Trust Factor) = $\sum Sv + Gv + SLAv + Av$

The complete process and all required inputs of OTF generation can be explained through Figure

IV. PREDICTIVE MODELING:

This section explores the potential use of predictive modeling in addressing the trust concern of the consumer in the provider. Predictive modeling uses variety of techniques e.g. data mining, machine learning and statistical analysis to predict the future measures based on perceived events. This paper will make use of Decision Tree, a supervised data mining algorithm for making predictions. Decision tree is a classification based algorithm that accepts both numerical and categorical data. It can generate rules that are easy to understand. One of the advantages of the decision tree is that it also indicates the importance of fields.

Figure 6 : Execution of Classification algorithm of Trust Prediction



1.8. Data Source: Decision tree algorithm was applied on the data set that consists of provider's status about recommended certifications and trust value. The trust value for the provider was calculated on the basis of proposed model explained in point 3.6 and 3.7 [Reference Table 3]. Decision tree is supervised algorithm; each supervised algorithm has dependent and Independent variables. In the current scenario [Security, Governance, SLA, Audit] were considered as independent variables. Trust is an dependent variable here [Reference Figure 6]. All the values are consumed in construction of the Decision Tree.

Security	Governance	SLA	Audit	Trust
2	0	0	1	Low
1	1	0	0	Medium
1	0	0	1	Medium
1	0	0	1	Medium
1	0	0	1	Low
1	0	0	0	Medium
1	1	0	0	Medium
0	0	0	0	High
1	0	0	1	Medium
1	0	0	2	High

1.9. Tree construction and Prediction: Lot of algorithms can be used to generate decision tree, current research is using ID3 method of generating decision tree. ID# was introduced by J.R.Quinlan. Quinlan has used entropy and Information Gain in construction of Decision Tree. The complete dataset is divided into training and testing data. The standard format 7:3 is used for division. 70% data is used for training and rest 30% is used for testing. Post division training dataset was used to create tree.

1.9.1. Entropy: Entropy is the measure of the randomness in the information being processed. An entropy method of calculating diversity index or homogeneity of the sample. If the sample is homogeneous entropy is 0 and if the sample is equally divided entropy is 1.

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

1.9.2. **Information Gain:** The information gain is based on the decrease in entropy after a dataset is split on an attribute. The best splitter is the one that decreases the diversity, so the attribute with the highest index (information gain) has to be the splitting attribute.

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)

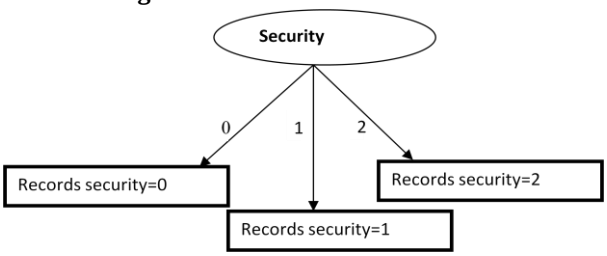
Entropy (l)= -m/ln * log2(m/ln) – n/ln * Log2(n/ln)

Calculating Information Gain of feature from Equation 1 to 5

Information Gain(Predictor_Variable): = Entropy (E) – j/N * Entropy (j) – k/N * Entropy(k) – l/N * Entropy(k) ……..Equation 5

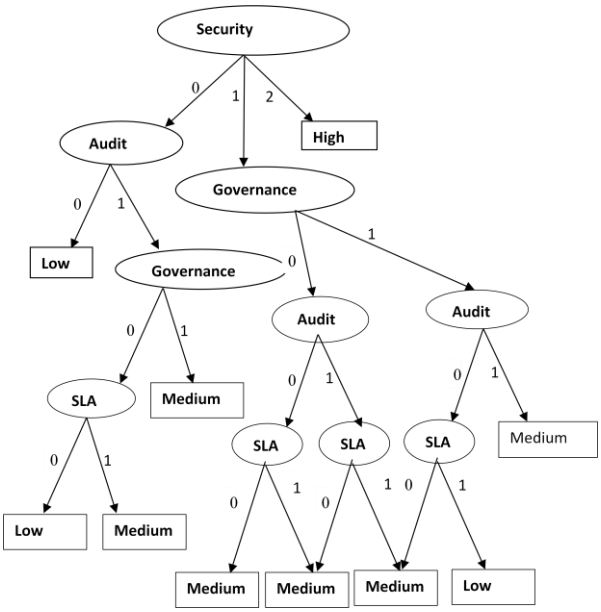
The attribute with the highest information gain will be considered as the root attribute. Split is craeted in the tree on behalf of highest information gain variable. Then the branches of the tree will contain records on the basis of split. In our case security has the highest information gain. The current of the tree can be viewed in Figure 7]. Since security has three values, there will be three branches from root node. Now data set will be divided in three parts:

Figure 7 : Level 0 Decision Tree



If all the records of the target variable has same values means we have pure dataset then that value is considered as the leaf node otherwise further a split is created in the tree. Same method of information gain is used for creating each split. The final structure of the tree after consuming all independent variables looks like Figure 8. Decision rules are extracted from decision tree.

Figure 8: Final Structure of Decision Tree



1.10. Result Analysis: After predicting for the unknown dataset, it is required to check the accuracy of the algorithms and predictions made by it. This research has adopted the following methods to check the outputs generated by the proposed model or classifier

1.10.1. Accuracy Score: Accuracy score indicates how accurate your classifier is or robust your classifier is. Accuracy explains, how many instances it classifies correctly and robustness explains, it should not miss the significant number of instances.

1.10.2. **Confusion Matrix:** A confusion matrix is a table that is often used to describe the performance of the classifier on the set of test data for which the class label or true values are known.

Here the classifier is Decision Tree

Table 4 confusion Matrix for Decision Tree (ID3)

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

V. CONCLUSION AND FUTURE SCOPE:

Trust issue can be resolved to certain extend if consumer can evaluate the trustworthiness of cloud provider before enrolling for the service. Current piece of research is supporting the cloud consumer in evaluation of trust value through information displayed in the public domain of the provider. The author has proved the efficiency of current algorithm through predictive modeling technique. One of the supervised learning techniques was used in this paper, there are many more that can evaluate the competence of the proposed model. The next step in this series is to develop a web based application that can compute trust value based on proposed algorithm discussed in the [3.6 and 3.7] points of this paper.

REFERENCES:

1. "Cloud Data Breach statistics." *https://www.hipaajournal.com.* n.d. https://www.hipaajournal.com/healthcare-data-breach-statistics/ (accessed 7 15, 2019).
2. *Data Breach In The Cloud – 2018 Trends That IT Pros Must Think.* n.d. https://www.cloudcodes.com/blog/data-breach-in-the-cloud.html (accessed November 27, 2018).
3. *Data breach incidents in India higher than global average* . july 23, 2018. https://economictimes.indiatimes.com/tech/internet/data-breach-incidents-in-india-higher-than-global-average/articleshow/65107118.cms (accessed March 11th, 2019).
4. RICHMAN, DAN. "Cloud Revenue." *https://www.geekwire.com.* 9 2, 2016. https://www.geekwire.com/2016/charts-cloud-computing-industry-getting-huge-decimating-sales-premise-servers/ (accessed 7 15, 2019).
5. Saxena, Archana B, and Meenu Dawe. "Certification Attainment - A Gizmo to Evaluate Provider's Trust." *International Journal of Natural Computing Research*, 2019.
6. —. "IAAS Service in the Public Domain: Impact of Various Security Components on Trust." *Information and Communication Technology for Sustainable Development.* Singapore: Springer, 2019. 789-797.
7. —. "Loss of trust at IAAS level: Causing factors & mitigation techniques." *Computing and Communication Technologies for Smart Nation (IC3TSN), International Conference on.* Gurgaon, India: IEEE, 2017.
8. Zhou, Minqi , Rong Zhang, Wei Xie, Weining Qian, and Aoying Zhou. "Security & Privacy in Cloud Computing : a survey." *https://ieeexplore.ieee.org.* November 3rd, 2018. https://ieeexplore.ieee.org/document/5663489 (accessed July 22nd, 2019).