SVM ALGORITHM FOR CYBER HACKING BREACH MODELING AND PREDICTION

Dr. DABBU MURALI, *Professor & Principal*, Department of Computer Science & Engineering,
VIVEKANANDA INSTITUTE OF TECHNOLOGY & SCIENCE, KARIMNAGAR, TS, INDIA.
E-mail: dabbumurali@gmail.com

ABSTRACT: Cyber incident data sets are essential for understanding the evolution of cyber threats. Since this is such a novel field of study, we have a lot to learn. This research includes a statistical analysis of data covering cyberespionage activities, principally malware infections, over a 12-year period (2005-2017). We find that random processes, as opposed to autocorrelation-based distributions, better explain the frequency and severity of cyber breaches than previous studies have revealed. Then, we provide our own special models of stochastic processes that are tailored to the sizes of the breaches and the times between their occurrences. We further demonstrate that these models can accurately foretell both arrival times and damage amounts. To better understand the evolution of hacking incidents over time, we conduct qualitative and quantitative trend analyses on the dataset. There are a few things that may be mentioned regarding cyber security. One is that while the total number of hacks has remained stable, the frequency with which they occur has increased.

Keywords: *Analysis cyber incidents, stochastic process, prediction of hacking.*

1. INTRODUCTION

Information theft is among the worst possible outcomes of an online interaction. According to the Privacy Rights Clearinghouse, there were 7,730 instances of unauthorized data access between 2005 and 2017, resulting in the loss of 9,919,228,821 records. In 2016, 1,093 data leaks were discovered, according to the Identity Theft Resource Center and Cyber Scout. Since 2015, when there were 780 breaches, that figure has increased by 40%. According to the US Office of Personnel Management (OPM), 4.2 million current and past Federal government employees had their personal information compromised in 2015. The federal employee and contractor databases also had background check information taken. In total, 21.5 million SSNs were contained within these files. Data breaches might result in significant financial losses. In 2016, IBM found that for every record containing private or sensitive information that was lost or misused, the cost was $158, on average. According to NetDiligence, in 2016 a median of 1,339 records were stolen at a median cost of $39.82 each. In the worst case, it cost $665,000. The cheapest one was $60,000. Data breaches are still a huge problem, even though technical solutions can help safeguard computer systems from attackers. This means that we must demonstrate the development of data leaks. This research will shed light not just on the prevalence of data breaches but also on other means of mitigating their effects, such as insurance. Developing reliable cyber risk indicators for allocating insurance premiums is complicated by our current understanding of data breaches and the absence of modeling techniques. Despite widespread agreement about the value of insurance, this situation persists. Recently, scientists have begun to simulate data-hacking attacks. Statistics on identity theft in the United States were analyzed for the years 2000 through 2008. There was a sharp increase in vacation time between July of 2000 and July of 2006. Then they began to stabilize.

2.LITERATURE SURVEY

Authors of the paper include Hammouchi et al. Al developed a social media–integrated STRisk predictive system to broaden the applicability of the prediction task. Approximately 3,800 businesses across the US were investigated. Both victim and non-victim populations were included in this. A profile is built for each company, with both technical indicators verified by external sources and social elements. The researchers acknowledge the possibility of errors in the non-victim sample and propose a solution for correcting incorrectly named organizations in order to address the issue of unreported cases. Next, they construct a variety of machine learning models to foretell the ease with which hackers could breach a certain organization's defenses. By considering both technical and social aspects, they are able to achieve an AUC of greater than 98%. Compared to the AUC they obtained while considering solely technical aspects, this result is 11% better.

The most reliable technical indicators are open ports and expired certifications, according to our research. However, the most reliable measures of social success are people's agreeableness and their capacity to share information. Mandal and the gang. Everyone participated in an effort to improve the social mood categories by combining different aspects of social events, responses, and their interrelationships. The proposed system not only plans ahead and notifies users of upcoming socially relevant events, but it also manages major social events. The obtained text data from Twitter datasets was subjected to an aspect-based mood analysis. It has been shown to be superior to cutting-edge methods.

Poyraz and the rest of his staff Costs associated with data breaches are examined from a variety of angles. In this article, we'll show you how to estimate how much money you'll lose due to a data breach. The system uses a dataset acquired from multiple sources that organizes stolen information about people living in the US into two groups: personally identifiable information (PII) and sensitive personally identified information (SPII). They use a sophisticated stepwise regression method that evaluates the influence of the independent components on numerous levels utilizing polynomial and factorial algebra. There have been three major breakthroughs in this field. To begin, our model establishes a correlation between the total cost of data breaches and factors including revenue, the quantity of PII and SPII exposed, and the likelihood of class action litigation. In addition, splitting secret information into "sensitive" and "non-sensitive" categories gives a more thorough picture of the expenses compared to past research. At the end of the day, there are many factorial interactions between the various independent variables.

From 2005-2018, Guru Akhil and colleagues studied datasets documenting incidents of digital espionage with a particular emphasis on breaches. They demonstrate that stochastic cycles, not distributions, are more appropriate for demonstrating the frequency, timing, and severity of hacking attacks, which runs counter to the findings of the study. Reason being, cycles point to inward dependencies. Next, they advise trying out multiple stochastic cycle models to locate the one that best describes the connection between the entry time and the extent of the break. These models have successfully predicted when and how large breaches between 21 will occur. They conduct subjective and quantitative pattern analyses on the data to learn more about how cyber intrusions occur as early as possible. Data on network security provides ample evidence to support the conclusion that cyberattacks are becoming less common but more dangerous.

The gang of Fang We started an investigation into how to model and foresee the danger of data breach at the enterprise level Due to insufficient training data, standard statistical models are immediately rejected when only a small number of breaches have occurred at a given company over time. They propose a novel and cutting-edge statistical strategy that makes use of the interconnectedness of various time series as a first step toward a solution.

They put the strategy through its paces by using a data set of actual enterprise-level hacks as test cases. The investigation clearly demonstrates its ability to model and predict company breaches. The gang of Kure. It depends on how significant assets are, how well present measures perform, and how well risk categories are forecasted in order for cyber security risk management (CSRM) to fulfill all of its aims. The proposed unified response draws from a wide range of techniques, such as fuzzy set theory for valuing assets, machine learning classifiers for anticipating potential dangers, and a comprehensive assessment model (CAM) for evaluating the efficacy of current safeguards. In order to estimate the level of danger, it is recommended to examine the relationships between key concepts from CSRM, such as asset, threat agent, attack pattern, strategy, technique, and procedure (TTP), and various subsets of the VERIS community dataset (VCDB). The experiment's findings suggest that applying fuzzy set theory to determine an item's significance enhances risk management across the board. The findings also demonstrate the efficacy of machine learning models in identifying various forms of risk, such as those posed by cyber espionage, crimeware, and denial of service attacks. An accurate risk forecast can help organizations figure out the best strategies to mitigate problems before they materialize.

Authors of the study are listed as Subramanian, et al. A model powered by machine learning was developed to fortify a website's defenses against hacking attempts. The primary objective of this work is to develop a machine learning model that can keep tabs on a website or system in real time, learn from its data, and adapt to the presence of new attack patterns. A web program built with Django and pulling data from multiple sources including Amazon, Flipkart, Snapdeal, and Shop Clues is proposed as a solution. This program securely displays the obtained data online. The proposed framework will monitor the location continuously, around the clock. We will save the data we collect in a way that makes sense and in a secure location on our platform, so that unauthorized third parties can't access it. The algorithm is updated daily and its estimates are based on state-of-the-art attacks and publicly available statistics. This model will be trained utilizing both datasets that already exist and records of assaults and breaches on our website from the past.

3.PROPOSED SYSTEM

We bring three distinct insights to bear on this issue. To begin, we demonstrate that the average time between cyber breach incidences and the amounts of breaches are better described by stochastic processes than distributions. How often something occurs can be inferred from the median time between occurrences. We discovered that the intervals between hacking breaches may be exhibited by a certain point method. Furthermore, we have discovered that a specific ARMA-GARCH model is effective at depicting the time series of hacker breach sizes. Here, ARMA refers to "Auto Regressive and Moving Average," while GARCH stands for "Generalized Auto Regressive and Conditional Heteroskedasticity." To the best of our knowledge, this is the first study to demonstrate that these aspects of cyber threats are better modeled as stochastic processes rather than distributions. Furthermore, we demonstrate that the number of events and the times they occur are directly related, and that this relationship is accurately captured by a certain copula. Furthermore, we demonstrate why it is crucial to factor in dependencies when calculating intervals between arrivals and the sizes of gaps between defenses. Failure to account for this will lead to inaccurate forecasts. To the best of our knowledge, this is the first attempt to investigate whether or not such a link exists, and what might happen if we choose to ignore it. We meticulously employ both qualitative and quantitative research methods to identify and analyze cyberattack patterns. The duration between cyber breaches is decreasing at the same rate as the average size of each breach remains constant. It's likely that the damage caused by each breach won't increase significantly, even if the number of breaches increases.

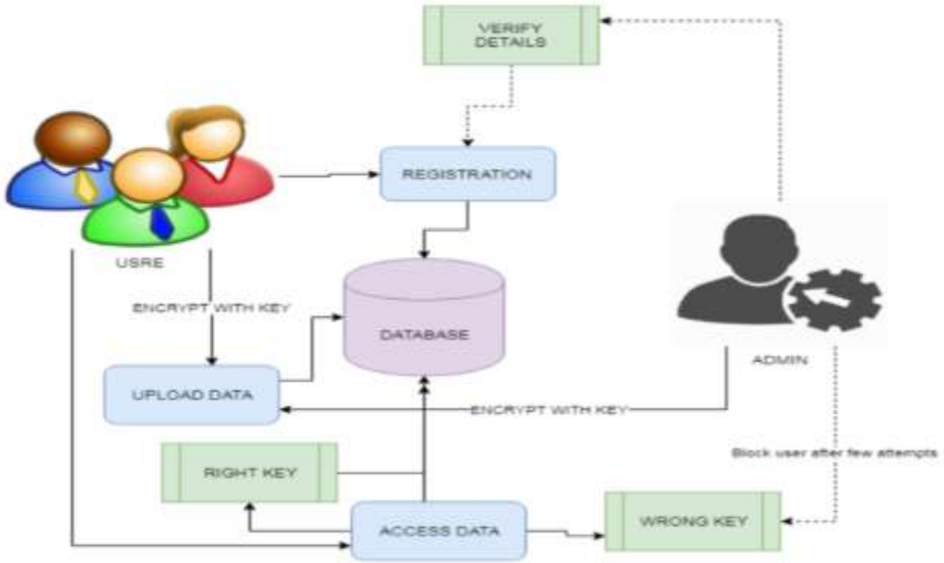


Fig. 1: block diagram of proposed system.

It is our sincere desire that this study would encourage more investigation, which in turn will yield significant novel insights into further means of reducing hazards. Insurance firms, government organizations, and lawmakers can all benefit from a better understanding of the complexity of data breach risks.

Support Vector Machine

The "Support Vector Machine" (SVM) is a supervised machine learning technique used to address classification and regression issues. Nonetheless, it is a common term in discussions of categorization. Using this technique, the n-dimensional space represents the total number of features, and each data point represents a single feature. Each characteristic's value corresponds to a location in this matrix. Finding the hyperplane that divides the data into two equal halves is the next step (see diagram below). The observation data are utilized to calculate the support vectors. A support vector machine (SVM) is a mathematical model for performing tasks like sorting, regression, and finding outliers by constructing hyperplanes in a space with many dimensions or an infinite number of dimensions. Separating the classes can be done effectively by using the functional margin, which is the hyperplane farthest from the nearest training-data point in any class. The rationale for this is that a smaller generalization error by the classifier is associated with a larger margin. Even though the first task may be performed in a very constrained environment, distinguishing between the various sets of interest is not always as simple as drawing a straight line between them. In order to emphasize the contrast more, it was proposed that the original space, which only had a few dimensions, be transformed into a much bigger space.

4. RESULTS MODULES UPLOAD DATA

Data resources can also be added by users who have been granted permission to do so by the database administrator. Encryption can be used to prevent unauthorized parties from gaining access to the information. Information provided to the administrator is used to determine authorization, and the administrator alone can provide authorization. Files can only be uploaded and downloaded by verified users.

Access Details

Administrators are in charge of allowing people access to the database's info. The administrator is the sole person who can manage user accounts and select who is granted access based on their credentials. The administrator is responsible for monitoring the information exchanged.

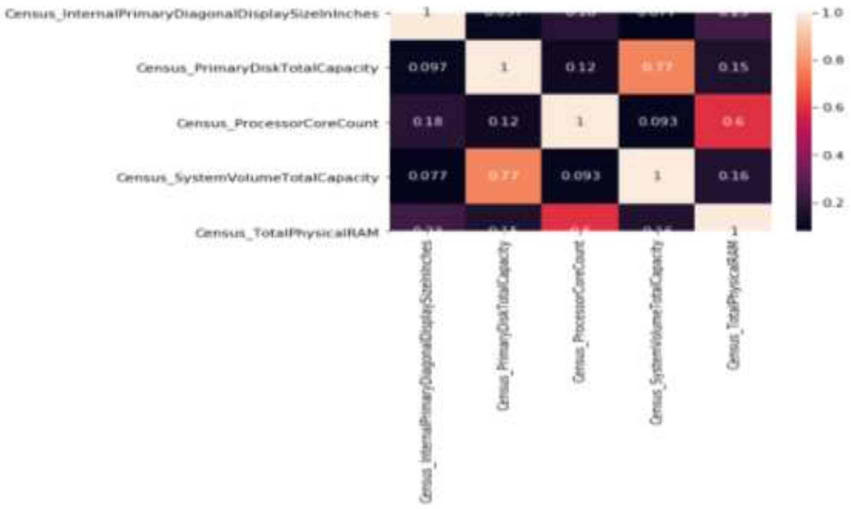
User Permissions

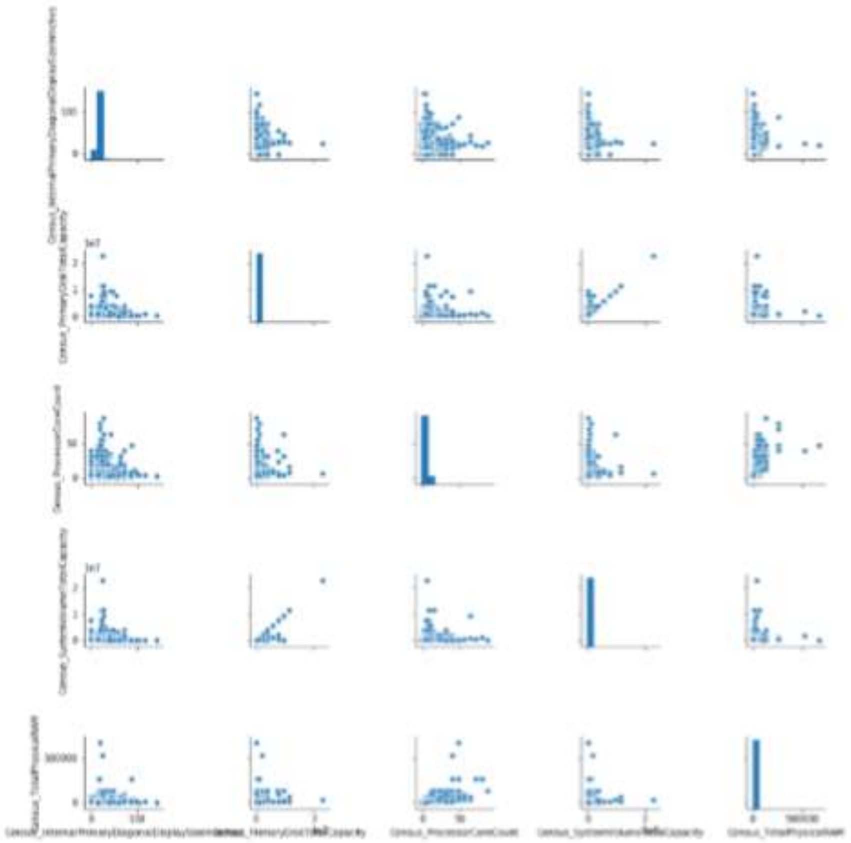
Data from any specific resource is hidden from everyone except administrators. The administrator notifies the user that they may share the information and then verifies the user's identity. If a user fails multiple times while trying to access the same piece of data, that user will be permanently blocked from doing so. The manager will consider the user's requests and past actions before deciding whether or not to unblock them.

Data Analysis

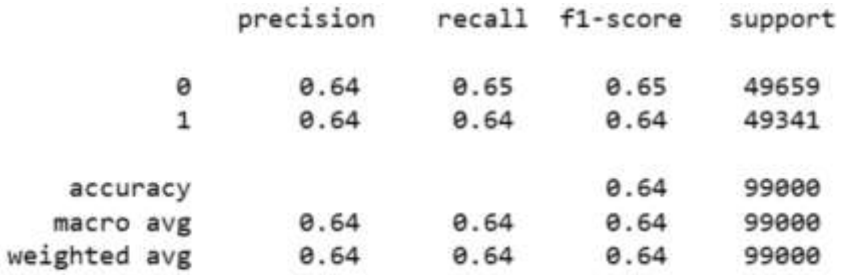
The use of graphs in data analysis is widespread. The collected information is visualized so that it can be studied and predicted with the highest precision and in accordance with the data laws. The data's visual representation can be used to aid readers in comprehending its contents.

EDA results





Classification report

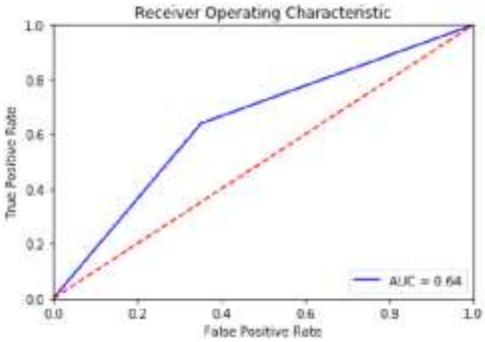


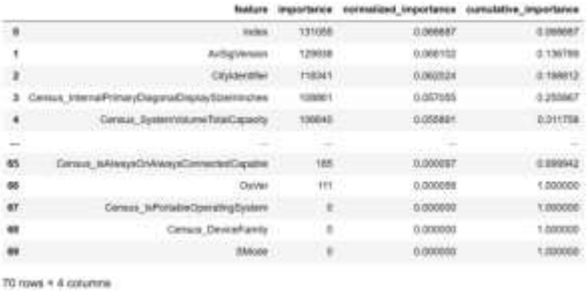
```
              precision    recall  f1-score   support

           0       0.64      0.65      0.65     49659
           1       0.64      0.64      0.64     49341

    accuracy                           0.64     99000
   macro avg       0.64      0.64      0.64     99000
weighted avg       0.64      0.64      0.64     99000
```

Feature Importance





	feature	importance	normalized_importance	cumulative_importance
0	index	131056	0.066687	0.066687
1	AvSigVersion	129938	0.066102	0.136789
2	CityIdentifier	118341	0.062024	0.198812
3	Census_InternalPrimaryDiagonalDisplaySizeInInches	109861	0.057055	0.253867
4	Census_SystemVolumeTotalCapacity	106640	0.055891	0.311758
...	...	...	...	...
65	Census_IsAlwaysOnAlwaysConnectedCapable	185	0.000097	0.999942
66	OsVer	111	0.000058	1.000000
67	Census_IsPortableOperatingSystem	0	0.000000	1.000000
68	Census_DeviceFamily	0	0.000000	1.000000
69	SMode	0	0.000000	1.000000

70 rows × 4 columns

6. CONCLUSION

We analyzed a database of hacking incidents to determine the average time between assaults and the volume of the disclosures. Our research suggests that, rather of using ranges to represent these two variables, we should instead rely on stochastic processes. When it comes to fitting data and making predictions, the statistical models utilized in this study excel. To calculate the overall probability of an impending occurrence with a certain violation magnitude degree, we advise employing a copula-based approach. In order to gain a deeper insight, tests were conducted using both qualitative and quantitative analysis. One of our research on cybersecurity reveals an increasing number of hacking attacks with a constant level of severity. These techniques can be modified and used to examine related data.

Future work

Some unanswered questions will require further investigation in the near and far futures. It's both fascinating and challenging to think about how to predict massive numbers and handle gaps in information like unreported security breaches. As an added bonus, knowing when leaks typically occur is a huge assistance. Also, more study needs to be done to look at how predictable breach incidents are, with the main goal of obtaining the best level of accuracy in forecasts.

REFERENCES

1. P. R. Clearinghouse. Privacy Rights Clearinghouse's Chronology of Data Breaches.
2. Accessed: Nov. 2017. [Online]. Available: https://www.privacyrights.org/data-breaches
3. ITR Center. Data Breaches Increase 40 Percent in 2016, Finds New Report From Identity Theft Resource Center and CyberScout. Accessed: Nov. 2017. [Online]. Available:

http://www.idtheftcenter.org/2016databreaches.html

4. C. R. Center. Cybersecurity Incidents. Accessed: Nov. 2017. [Online]. Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents
5. IBM Security. Accessed: Nov. 2017. [Online]. Available: https://www.ibm.com/security/data-breach/index.html
6. NetDiligence. The 2016 Cyber Claims Study. Accessed: Nov. 2017. [Online]. Available: https://netdiligence.com/wp-content/uploads/2016/10/P02_NetDiligence-2016-Cyber-Claims- Study-ONLINE.pdf
7. H. Hammouchi, N. Nejjari, G. Mezzour, M. Ghogho and H. Benbrahim, "STRisk: A Socio- Technical Approach to Assess Hacking Breaches Risk," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2022.3149208.
8. Mandal, S., Saha, B., Nag, R. (2020). Exploiting Aspect-Classified Sentiments for Cyber- Crime Analysis and Hack Prediction. In: Kar, N., Saha, A., Deb, S. (eds) Trends in Computational Intelligence, Security and Internet of Things. ICCISIoT 2020. Communications in Computer and Information Science, vol 1358. Springer, Cham. https://doi.org/10.1007/978-3-030-66763-4_18
9. Poyraz, O.I., Canan, M., McShane, M. et al. Cyber assets at risk: monetary impact of U.S. personally identifiable information mega data breaches. Geneva Pap Risk Insur Issues Pract 45, 616–638 (2020). https://doi.org/10.1057/s41288-020-00185-4
10. Guru Akhil, T., Pranay Krishna, Y., Gangireddy, C., Kumar, A.K. (2022). Cyber Hacking Breaches for Demonstrating and Forecasting. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_106
11. Z. Fang, M. Xu, S. Xu and T. Hu, "A Framework for Predicting Data Breach Risk: Leveraging Dependence to Cope With Sparsity," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2186-2201, 2021, doi: 10.1109/TIFS.2021.3051804.
12. Kure, H.I., Islam, S., Ghazanfar, M. et al. Asset criticality and risk prediction for an effective cybersecurity risk management of cyber-physical system. Neural Comput & Applic 34, 493–514 (2022). https://doi.org/10.1007/s00521-021-06400-0
13. R. R. Subramanian, R. Avula, P. S. Surya and B. Pranay, "Modeling and Predicting Cyber Hacking Breaches," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 288-293, doi: 10.1109/ICICCS51141.2021.9432175.