Data Science in Action: Key to Cybersecurity

Afreen Khan  
Research Scholar  
Department of Computer Science  
Aligarh Muslim University  
afreen.khan2k13@gmail.com

Prof. (Dr.) Mohammad Ubaidullah Bokhari  
Department of Computer Science  
Aligarh Muslim University  
mubokhari@gmail.com

Abstract - The coming of Big Data has completely shaken the computing world. With the rise in voluminous data, the amount of data at times denotes to the predictive capability of a particular data model and other times to the discovery of hidden insights that appear when rigorous analytical methods are applied to the data itself. Data Science is the application of advanced analytics to activity and access data to uncover unknown risks. It is the practice of deriving valuable insights from data. From a cybersecurity point of view, the value of data refers first to the nature of the data itself. The malicious network traffic data from malware and cyberattacks have much more value than some other data science problems. With respect to security, the valuable insight leads to reduced risk. Data Science is an emerging field that is proving well to deal with the challenges of handling alargeset of data and the outbreak of new data produced from theweb, mobile devices and social media. The IT sector that deals with the information security and prevention of frauds has been continuously advancing, using data science techniques to deal with the challenges of handling and achieving insights from vast amount of log data, uncover insider threats and attacks therein prevent the frauds. The data science methods used for cybersecurity applications is a moderately new concept. This comprises deployment of statistics, machine learning, mathematics and big data analytics for anomaly detection, network modeling, risk management and more. This paper, thus illustrates how the data science tools and practices can act as a boon in combating with the cybersecurity, what role the machine learning plays in securing the web and the benefits and application of using data science for cybersecurity.

Keywords - data science; cybersecurity; machine learning; big data; statistics; big data analytics

I. INTRODUCTION

A report by Databricks, 2017, stated that in the first half of the year 2017, 1.9 billion records have been stolen. It was noted then, that on an average the cost of a breach is about $3.62 million and more than 6 terabytes of data per month are being collected and analyzed by the security teams. As we are heading towards the era of 21st century, data deluge is at its peak. The more data is collected, the more sophisticated methodologies the scientists need to apply in order to gain insight from the collected data. In today’s data-driven world, organizations comprehend that there are severe threats that must be addressed. The complexity of IT networks has escalated rapidly in the past few years thus growing the sophistication of cybersecurity threats and attacks [1]. According to the latest statistics by PandaLabs, about one billion incidences related to malware occurred between June and November 2016 in which the average cost of a breach was about $1 billion and 99% of computer systems were under the cyber-attacks [2].

The Fig. 1 below depicts the total cost of cyber-crime in seven countries, namely, US, Germany, Japan, UK, Brazil, Australia and Russia for the fiscal years 2013, 2014 and 2015 as described by a global report by Ponemon in 2015.

![Fig. 1. The total cost of cyber-crime in 2015 in seven countries](image)

This suggests us that as the malware threats and attacks are increasing in quantity and complexity, conventional analytic
tools and techniques fail to make up for the huge volume of data and its scalability. To detect and prevent the insider threats has become more difficult to combat with. Therefore in order to eliminate, these insider attacks hinge on uncovering irregular user behaviour [3]. The predictive analytics, big data analytics and machine learning are aimed to be the solution to the cyber threats and attacks that will aid the organizations in solving their problems [4]. Because the data breaches are becoming more sophisticated with the technology at hand, our methodologies to combat with them is also becoming defensive in nature.

The application of machine learning, mathematics and statistics used so as to mine the knowledge and discover relationships and patterns from the collected data, known as Data Science, is a budding and promising technology in the present times [5]. It can aid in shifting from static models to the dynamic models that can describe the user behaviour based on various characteristics like identity, tasks and working conditions thereby facilitating in lessening the false positives and emphasizing the user behaviour that strictly accounts for malevolent actions [3]. Thus it can be stated that the data science is on its way to rise, empowering the cybersecurity sector to shift from hypothesis to facts thereby it can be proved to be effective in combating with the threat detection and prediction.

Researchers around the globe are emphasizing on building efficient tools and technologies to examine the fields of what constitutes Data Science in the domain of cybersecurity. Thus, this paper will include the evolving realm of Data Science in the age of big data analytics and machine learning. The paper is organized as follows: The subsequent section involves the state of big data and big data analytics. section 3 discusses defining data science, and role of machine learning in security. Section 4 introduces the challenges in countering the cybersecurity threats. Section 5 presents data science- as the path forward to cybersecurity, and applying machine learning to security. Section 6 introduces the application to cybersecurity followed by benefits and conclusion in Section 7 and Section 8.

II. THE STATE OF BIG DATA AND BIG DATA ANALYTICS

Big data- a universal deluge of digital information, is used to define the enormous quantity of data; the data being in unstructured, semi-structured and structured form. Big data is outlined with respect to the 3 V’s, namely, volume, variety and velocity respectively. The big data analytics comes into picture when the advanced analytic methods function on huge datasets. Therefore, big data analytics sums up to two things- big data and analytics. The big data tools and expertise possess the potential to transform our abilities to deal with the huge datasets produced in any cyber data analytics [6]. The matter of the fact is not just to handle the huge quantity of data, but to control all the collected data in order to offer enhanced and quicker analytics for threat and attack detection [6].

Big data analytics encompasses cutting-edge analytics along with the numerous additional technologies that are functioned on the big data which delivers significant information. Big data analytics methodologies blend together machine learning, text mining and ontology to lessen and improve the security threat prediction, uncovering and avoidance at a primary phase [7]. The 3 V’s of big data are all vital features that possess high significance for using data science to cybersecurity [8]. As the gap for data rises, the requirements for computational processing also increases so that complex and severe workloads can be handled accordingly [9].

III. DEFINING DATA SCIENCE

The notion of mining information from data is not a new concept, but the idea of data science is a self-sufficient and unconventional field of study [10]. Data science encompasses concepts from mathematics, machine learning, statistics and big data analytics. Generally, the data, in its raw form is gathered, processed, analyzed, and interpreted using different algorithms and models and the outcome received is then used to deduce the recommendations and delivering the final product, known as a data product. This field of science embraces a broader variety of analytical methods and applications, thereby acting as a boon for the organizations in handling various world problems along with the concept of big data and analytics.

The central idea behind the growing fame of data science is the tremendous volume of digital data being produced, stored and processed every day. To counteract with cyber attacks, various data science methodologies and advanced analytic techniques are being adopted for the real-time data so as to derive insights from the vast quantity of data produced thus posing with demanding challenges while reducing both false positives i.e. false alarms and false negatives i.e. failing to identify existent threats [8].

A. Role of Machine Learning in Security

Machine learning practices come up with numerous novel statistical, mathematical and scalable challenges [6]. As the data is growing at a huge rate, the organizations are confronted with the complex responsibilities and hence, to scale such a large amount of data becomes a critical concern in these systems. Thus, such a heap of experimental data can be processed and analyzed using machine learning effectively [11]. On a broader level, the machine learning is being used to locate the anomalies. Specifically, it is used to detect malevolent behaviour or threats such as attackers, hackers, malware, redundant behaviour etc [12]. The machine learning algorithms give an advanced insight into big data which thereby can be optimized and used in a probable way to further improve the related performance metrics in the various applications. These algorithms, tools and practices differ to a great extent in terms of distinctive functions. For example,
Data Science in Action: Key to Cyber Security

support vector machines, decision trees, advanced clustering and deep learning.

To analyze and process this huge amount of data, machine learning and cybersecurity are suited well that when applied can generate an evolutionary method and thus continuous production of such algorithms can give a better optimization [11]. The threats and attacks on platforms and networks are continuous and are much efficient and successful given the set of tools to examine and assess the targets and hence, therefore, the opponents today are applying machine learning methods so as to combat with these attacks [13].

IV. CHALLENGES IN COUNTERACTING THE CYBERSECURITY THREATS

The 21st century, technologies at hand, ease of use, possess a great demand for the cybersecurity of the data being gathered every day. There are four severe challenges that need to be looked upon in order to deal with the cybersecurity threats and attacks, as defined below in Table I [14].

TABLE I. KEY CHALLENGES TO COMBAT THE CYBERSECURITY THREATS

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rising data volume</td>
<td>Every day a huge amount of data is produced all over the world from the various sources in which large set of files are reported for malicious data which thereby requires processing and analysis. As data size grows, a number of false positives generated by SIEM (Security Information and Event Management) tools also increases.</td>
</tr>
<tr>
<td>2. Failure to scale up</td>
<td>The conventional detection methods were not designed for terabytes of data. As the data volume grows, the performance metrics degrade and to scale up the storage is hence a costly task.</td>
</tr>
<tr>
<td>3. Absence of enhanced analytics capabilities</td>
<td>SIEM (Security Information and Event Management) tools and technologies depend on definite guidelines to aware teams of possible risks and threats. It crashes to recognize threats that are not defined by the guidelines.</td>
</tr>
<tr>
<td>4. Difficult and inflexible to use</td>
<td>The conventional security tools and practices are rigid in their approach. The SIEM tools enforce copyrighted data formats and query languages but the majority of the tools don’t possess real-time collaboration with others.</td>
</tr>
</tbody>
</table>

V. DATA SCIENCE: A PATH FORWARD TO CYBERSECURITY

During the past 15 years, numerous analytic tools and techniques have been extended to leverage the scaling model of big data in order to tackle with the log data and the linked-node data [6]. The cybersecurity organizations have adopted such methods and have increased to use this technology to combat the insider threats and attacks [3]. The data science is a path that has shifted its core working from static models to the dynamic models, which combines the data processing, domain expertise, machine learning, mathematics, statistical research and computer science so as to deliver the data product.

The cybersecurity team leverages the log data and linked-node data of social sites in order to manage the huge volume of data so that the pattern of the particular sessions can be determined and how much appropriate the activity is in between the resources can also be controlled [6]. This employs a fusion of data science tools and techniques, domain experts, security experts and threat intelligence to provide correct and actionable data with reference to the security alarms and discover anomalies and indications of insider threats.

NIST (National Institute of Standards and Technology) has built a framework that incorporates cybersecurity and data science, known as “Framework for Critical Infrastructure Technology” [10]. It includes five key objectives as defined below in Table II:

TABLE II. FIVE KEY OBJECTIVES FOR THE “FRAMEWORK FOR CRITICAL INFRASTRUCTURE TECHNOLOGY”

<table>
<thead>
<tr>
<th>Objective</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identify threat assets and the aftereffects of the breach</td>
<td>The assets are ranked according to the possibility of an attack and the consequences of the breach. In this way, a better behavioral quantitative model can be developed.</td>
</tr>
<tr>
<td>2. Protect the information</td>
<td>In this, the patterns are identified first and then algorithms and definite signatures are developed to prevent and safeguard the information.</td>
</tr>
<tr>
<td>3. Detect the intruders</td>
<td>A rule-based expert system is built that can detect problems inside the network system and accordingly generate a response.</td>
</tr>
<tr>
<td>4. Respond to the intruders</td>
<td>The data scientists can make use of this gathered response in order to analyze the efficiency of the response and simultaneously the source of the defect too.</td>
</tr>
<tr>
<td>5. Recover the business</td>
<td>The business continuity and its resilience are the key features of any data breach. Hence the data scientists can be engaged so as to deal with these key aspects in order to recover the business.</td>
</tr>
</tbody>
</table>

A. Applying Machine Learning To Security

In machine learning lingo, from a supervised perspective, there can be large sets of labelled data and missing labelled data. The former category is used in malware detection and spam detection whereas the latter being used in anomaly detection and risk scoring. The unsupervised side looks a bit different. It consists of clustering, association rule learning and dimensionality reduction. The first two are used in entity classification, anomaly detection and data exploration [12]. These machine learning algorithms act as a boon in detecting the threats and the attacks when applied in the right manner. But these algorithms don’t prove well when a scientist need to identify the anomalies because they are meant for limited use. To deal with the malicious and anomalous activities, the
A majority of network protection mechanisms work well that gives log and other specific information but these logs are of less priority for a large number of security tasks and operations [13].

SIEM (Security Incident and Event Management) system is used to emphasize on the activities for the cybersecurity experts in which the logs and related information is fed into this system for the management of log [13]. But the cutting-edge threats hide their trail and take probable action to avoid the processing of log management. In a report of 2015 by McAfee, it has stated about 500 million new kinds of malware data [15]. The complexity of such malware creates identifying more cutting-edge and advanced malware that sound more complex and challenging for cybersecurity specialists which thereby puts a question mark for the detection and signaling of the malicious activities [15].

It is the machine learning which along with other data science methods and practices deal with the malicious activities in order to provide security react to cyber threats and attacks appropriately [13].

VI. APPLICATION TO CYBERSECURITY

In 2015, ‘US Government Office of Personnel’ was attacked and private information of about 21.5 million individuals was stolen in a colossal data breach. In 2016, the Bangladesh Bank’s systems were attacked and hackers stole the identifications for transferring the payments [16]. To counteract the cyber attacks, following questions need to be answered before implementing the suitable data science tools and technologies [6]:

1. If there is any data that is available for a particular malware that attacks universally?
2. The number of machines the malware affected and which of them compromised?
3. How many and what users were affected?
4. What ports were influenced?
5. What data was leaked?
6. If the sensitive information was lost, then who did- an outsider or an insider?
7. Globally, where did this particular malware has attacked before- all the exact location list?
8. Based on the past records of this attacker, what must one expect within the next hour or next week or next month from this attacker?
9. What insecure activities the particular organization’s users adopt to combat the attacks? When, where and what suspicious event happened and where is the maximum threat to the organization?

Using the appropriate statistical approach, susceptibilities versus attacks can be tabulated and the results can be visualized then accordingly. Keeping in view the respective outcome, a probable solution can be made in order to combat the threats using the proper data science practices. Furthermore, one needs to detect anomalies in the behaviour of a device, identify the anomalies in the behaviour of an employee and service provider, identify irregularities in the network system and evaluate network susceptibilities and risks [17].

Many organizations are on the list that has developed a solution to combat the cybersecurity threats and attacks. Some of them are described as follows:

1. Sophos, about 30 years ago developed encryption and antivirus products using big data analytics which now provides a service to about 100 million individuals and 1 lakh businesses across 150 countries to detect the malware and secure the networks. It includes: malware examination and its analysis, large-scale trend analysis and evaluating detection performance [17].

2. RDS (Rapid Detection Service) is the latest security method employed by the Finnish firms which provide a service of protecting against both outsider and insider attacks. RDS gathers data from varied sources and using the analytical methods detect when a particular user account perform abnormally.

3. The IBM Security methods have developed a system that analyses the data at risk, protect confidential data from internal and external events and speedily adjust to changes environmental changes.

VII. BENEFITS

Just detecting the possible threats and risks is not the real game. A probable solution is that which includes the 3 chief objectives- Prevent, Detect and Respond (PDF) [17]. The advantage of PDF is that using the operationalization skills, it can examine the gathered data, discover the correct signals and can then generate the exact actions [17]. Whenever the possible risk appears, it spontaneously freezes the devices that are located to steal the data, the machines receive an alert about who, when and where stole the suspicious data and send signals to the exact person in concern [17]. Another benefit of using the data science practices for cybersecurity is that specific risks can be geared up when specific tools to collect, process, analyze and share the data is applied. The chief objective must be to strengthen the cybersecurity community that should include skilled data scientists and big data analytics experts [19].

When the talk is about the big data analytics and data science practices, the more the data one has the better the analytical solution one obtain. Thus, given the large volume of data, threats and the malware attacks, it gives the promise of dealing with them using novel and noteworthy methods to crack the complex challenges and difficulties [4].
VIII. CONCLUSION

Data science tools and practices used for cybersecurity is a new paradigm. It comprises deployment of machine learning, statistical methodology, mathematics, big data analytics for anomaly detection, risk management, network modelling, and more. It is in a growing stage and is proving to be an essential component of cybersecurity. The objective must be to reduce the risk and shield the sensitive data. A combination of right individuals and right training on the team can leverage to build robust defence mechanisms against the cyber threats and hence can protect the data and business from damages [10]. It thereby aids in intelligence and investigation practices via better collaboration among the team so that the organizations can discover and detect the threats easily [7]. Organizations are now shifting their usage of traditional methods to the unconventional tools and practices. Data science have proved to be one such unconventional method that speeds up the detection process, thereby blocking the concerted threats and attacks immediately before it could further create damage.

REFERENCES


