Machine Learning assisted Shoe Size Recognition Model:Security and ethical issues

Heinrich Davids
Department of Economic and Management Sciences
University of the Western Cape
Cape Town, South Africa
2330838@myuwc.ac.za

***Abstract*—Technology has been the cause of hopes, fears and ethical consideration for many years now. The ubiquitous nature of technology makes it almost impossible to engage with it in some form or another. With the arrival of Big Data, the positive and negative connotations of technology have increased. Big Data can be identified by the four V's, Volume, Variety, Veracity and Velocity. Many researchers claim there are plenty more V's or characteristics, but these four are the most prevalent in characterizing Big Data. Organisations are increasingly benefiting from Big Data innovations, whilst equally battling to prevent data breaches and using data in an ethical manner. This research aims to identify the most common security and ethical issues organistions face, and how these issues are applicable to installing a Shoe Size Recognition camera in a fashion retailer store. The purpose if the camera is too identify shoe sizes of customers walking into the store, and use the data collected to improve on demand forecasting.**

Keywords—Big Data, Technology, Organisation, Environmental, Security, Ethics, Privacy, Choice, Trust, Awareness, Machine Learning, Analytics, Hadoop Distributed File System

I. Introduction

The world today is overwhelmed by data and it is increasing day by day [1]. Big Data (BD) is believed to be the new salvation for businesses to reinvigorate their competitive position in the market [2]. BD drives economic growth through technological innovation in businesses [2]. BD refers to data in large volumes, which are continuously generated from various sources (sensors, human beings, machines, equipment), and from multiple environments [3]. Big Data can be identified by the four V's, Volume, Variety, Veracity and Velocity. Many researchers claim there are plenty more V's or characteristics, but these four are the most prevalent in characterizing Big Data. Volume refers to the large size (Gigabytes to Petabytes) of BD; Variety refers to the multiple sources where the data is coming from, for example Social Media platforms, Video, Audio, sensors and ERP systems; Veracity refers to the unpredictable nature of BD, sometimes the data will be accurate and relevant and other times it can be inaccurate and irrelevant; Velocity refers to the speed at which the data is generated and collected.

It is the ubiquitous nature of BD that is giving it a bitter sweet connotation among society's members and business professionals. Every click on the internet or movement with a smart device is in the hands of internet business companies, like Amazon, Google and Facebook [2]. Business professionals are monitored more closely to track their performance and measure their KPIs, for example, truck drivers are monitored via Intelligent Transportation Systems. To make it worse, many of the above mentioned companies has had some form of data breaches in recent years, with many customers' information been leaked to malicious entities [2].

The fashion retail industry has traditionally relied on creativity, intuition and historic Point of Sale (POS) data for designing, buying and merchandising [4]. BD analytics can be used for a number of purposes inter alia, trend analysis, market identification, measuring influencers' impact and understanding the customer[4]. In a time of global uncertainty, with COVID-19 causing unpredictable markets and irregular spending patterns, BD can provide a much needed competitive advantage and improved profitability. Trend forecasting in the fashion industry offer great insight to planners, incorporating Google Trends technology. Google Trends can predict present and near future fashion trends with BD analytics [5]. Demand forecasting is as crucial in the fashion retail industry as it is in any. Knowing your customer's needs not only means the company sells it's products, but also eliminate overproduction which leads to production wastage and also reduce excess inventory which in turn leads to products being marked down and sold for much lower prices [4].

Historical POS data are deemed insufficient for accurate demand forecasting in the fashion retail industry [4]. Companies can only gather data on clients who actually made a purchase, non-purchasing customers are non-existent to retailers, which means they lose valuable information on their actual clientele. Tracking all feet that enters a shoe retailer can give insight to "all" customers walking in the store. Gender classification and shoe size identification by means of a Machine Learning (ML) algorithm, using the Convolution Neural Network technique to predict the gender and shoe sizes of customers walking into the store. This information can be invaluable to shoe retailers. The question is, is it too invasive to the customers, is it legal and what customer data will be saved by the company?

II. Literature Review

A. Security

Companies can split their security awareness and policies into three categories, Technological, Organisational and Environmental [6]. For this reason the Technology, Organisation and Environmental (TOE) framework will be applied to discuss Security considerations applicable to implementing a Shoe Size

Recognition Model (SSRM) in the fashion retail industry.

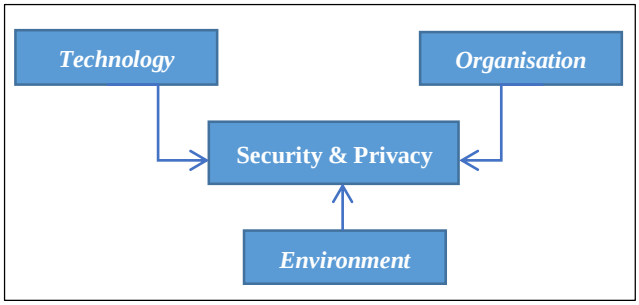


Figure.1 Adapted TOE framework

1) Technology

This section entails the organisation's internal equipment and processes. For BD a strong security solution are needed to ensure data integrity, confidentiality and availability [6]. Traditional security protocols used for traditional Relational Database Systems (RDBMS) for example are not sufficient in a BD environment [4]. BD is generated and analysed at unprecedented speeds, either in batch, real time or near-real time, and this makes it difficult to maintain data protection [6]. Another distinction from traditional data is the data security life cycle [9]. Traditional security technologies does not take into account the whole process of the data security life cycle, for example, data can be shared over multiple devices and cloud storing platforms, BD security need to consider the protection of the data across all platforms/ devices.

Another unique characteristic of BD that needs to be considered in terms of data security is the various formats and sources of the data being collected [6]. This variety characteristic of BD usually means the data is unstructured more than structured. Most security technology is not capable of protecting unstructured data [6]. With rapid transfer of data streams from local servers to BD platforms, for example cloud based storage; a cloud based firewall is needed for data protection and to prevent data loss [6]. Cloud based firewalls provide, availability, scalability and extensibility [7], which means backups in case of site failure, scalability to handle high bandwidth capabilities and a secure communication path. Cloud storage may pose security problems and lead to privacy issues when the data are hosted in a server that is publicly accessible, so companies needs to ensure sufficient vetting of cloud service providers.

Live video stream will be uploaded from an Internet Protocol (IP) camera to a cloud based platform directly from the camera for the video feed storage. The upload will happen via the store WiFi network. The store is part of the corporate brand Pepkor, which means there are already extensive data security protocols in place to ensure a secure upload connection, hence no extra security mechanisms will be needed in terms of the upload. Attention should be given when selecting the cloud storage service to ensure end to end protection and privacy of video footage of Shoe City customers. Cloud computing environments store data from different devices, so it can be a major security risk [11]. The Proof of Concept (POC) of the solution will be based on community based cloud service, and a thorough vetting process will have to be done before selecting a cloud service to ensure adequate protection is in place. Important data protection techniques to look out for are data confidentiality, integrity and data access controllability [11].

Video footage should be periodically removed from the cloud server as a precaution, once analysis has been done and the feedback from the SSRM has been uploaded to HDFS the video footage can be deleted. The cloud based storage service should also offer enough storage capacity and routine backups to prevent massive data loss.

2) Organisation

The Organisational aspect of the framework deals with company culture, strategies, structure and policies [7]. Even with the most advanced security protocols in place, if the staff does not adopt the culture of protecting company assets, the company is not protected. Organisational culture and awareness on security and privacy, is crucial in eliminating human-related security breaches [6]. In 2009 approximately 221 million personal identifying information records became public knowledge because of external information breaches [8]. A good example of human-related security breaches are through phishing [8]. Phishing is when an email user gets an email and the email user reveals confidential information such as usernames, passwords and bank account details by clicking on website link or follows certain instructions. This information can be used to infiltrate the company's networks and use the data for illicit purposes.

To avoid a catastrophic event of a data breach it is crucial companies have sufficient protection mechanisms in place [6], but for a successful security culture in a company technology is not enough, top management needs to promote the security culture and provide the necessary support to the staff [9]. A lack of top management involvement and support may deter all efforts made by IT professionals. Another important organisational security aspect when implementing BD is the level of employee competencies or understanding of BD protection mechanisms [3]. Adequate training should be provided to staff to ensure a good understanding of the company's policies and culture. Employees should acquire data privacy and security training through SETA programs as research suggests training is an effective method of expanding employees' knowledge [10]. A study done by [10], suggests prevented phishing attack incidents dropped a further 10% after employees were trained on the awareness of phishing attacks.

The in-store camera does not pose many data security risks in terms of organisational culture and policies, as all data will be uploaded to a remote cloud server. Adequate training should be provided to staff make sure they do not divulge the WiFi password to any unauthorised people and also to treat the camera company property and not tamper with it. If employees

put a memory card in the camera, they can get access to the data and this might pose an ethical problem when company data fall into unauthorised hands. The store manager will have access to dashboards that presents the results of the SSRM calculations, for better planning. This information is propriety to the fashion retailer and might not be divulged to any unauthorised people. Thus from an Organisational security perspective, sufficient training must be given and adequate awareness of data confidentiality and security should be created.

3) Environment

The environmental context refers to the environment where the company conducts its business in, which entails the industry, government and the industry [12]. The external environment always has an impact on the internal environment, and is harder to control. In the BD environment, the collection of data from numerous sources is typically involved and many times this data comes from other companies [13]. Data mining companies has become very popular over the last five years. These companies collect data on behalf of other companies, for example, companies that mine social media websites and sell customer interests to corporates. Once sensitive data are being shared cross organization, security, privacy and confidentiality issues arise [6]. Both companies at either side of the data transfer should have adequate security mechanisms in place, and a clear understanding about the intellectual property of the data should be established beforehand. Companies can use data in unethical or even illegal ways, so the data owners should be establish to prevent legal complications and also reputations being tarnished by the acts of other entities.

Integration between companies, like a retailer and their supplier, also poses a security risk as the companies exchange information or data between each other. Companies conceal their security profiles as they want to hide their vulnerabilities to protect their assets [14], so the companies won't necessarily know the security level implemented in the other company. The relationship between system integration and level of security countermeasures is dependent on the firm size, industry and other external factors [15]. In 2018 a study found, the greater the integration between companies, the larger investment companies make in security controls and countermeasures [14].

Another environmental aspect companies need to address when considering their BD security profile is the use of third party tools [6]. Usually in a BD environment companies use third party tools/ applications to store, analyse, access and share data [6]. For example, HDFS is prominently used for BD storage and processing within a BD environment [7]. Dependence on BD third party applications from companies using their services are prevalent because companies does not have the internal security mechanisms in place, thus they rely on third party's security infrastructure [6].

In terms of the SSRM, the data will not be integrated to external companies so no evident considerations need to be taken on that. The data will only be used for the company where the camera is installed in the store. The data will be uploaded to a HDFS, so security and privacy considerations are applicable in terms of the SSRM. Using the HDFS could be beneficial in terms of security mechanisms that do not currently exist within the client company. The data generated inside the retail store, which make the company the data owners of all the data and are responsible for ethical usage of the data.

B. Ethics

BD analytics makes use of algorithms to analyse huge datasets to detect patterns, similarities, and other economic and social value [16]. Many researches claims BD analytics has come under criticism recently for having unethical consequences to various stakeholders [17] [18]. Breaching of privacy, discrimination against customers and individual profiling are amongst the issues that has been raised [16]. Companies argue they collect data on consumers to deliver a more personalised and better quality service to the consumers, but consumers feel the incentives are more for the companies than for their needs [16]. The main BD ethical issues from an individual's perspective can be divided into four groups namely, Privacy, Trust, Awareness and Choice [16].

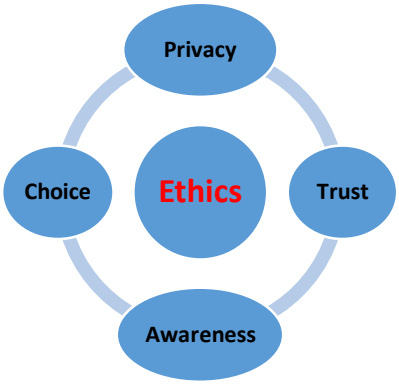


Figure.2 Privacy, Trust, Awareness and Choice

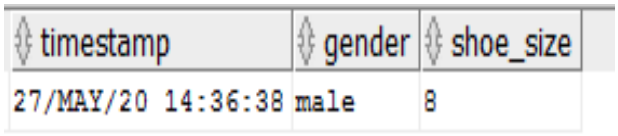
1) Privacy

Privacy in the context of BD ethics can be defines as the individuals' capacity to restrict and control how companies use their personal information [19]. This entails how companies access, modify and use personal data. Individuals feel even if they give consent to companies to collect and share their data, they want to decide what data are being collected. For example, if I give consent for companies to collect my Google search history, I still want to restrict my online purchasing data. The information given to companies can be filtered down to exclude personal information on individuals, and only show online searches or purchases and not demographic data, but the aggregation process might re-identify the individual data without the individual's knowledge [20].

Individuals also feel they should be able to modify the data about themselves, to avoid misrepresentation of themselves [16]. Individuals might make purchases on behalf of someone else then get inundated by online suggestions and unsolicited advertisements about those

products which are irrelevant to them. Another concern individuals have in terms of their privacy being violated is, sensitive data about them might get shared and create discomfort or make them victims of discrimination [16]. It is a known fact that certain government agencies "spy" on citizen's online activity, and might misread online activity of individuals.

Installing a video camera in the store to only collect data of their feet/ shoes is not invasive to the customer, but some customers might feel their consent is necessary to collect the data. For this a disclaimer needs to be displayed where the customer can see it. The disclaimer will count as consent given by the customer that the company may collect and use the data collected on the customers. No personal data on the customer will be collected, only their shoe size and their gender that will be added on to the SSRM at a later stage. Below an example of the data in the HDFS:



timestamp	gender	shoe_size
27/MAY/20 14:36:38	male	8

Figure.3 Example of data uploaded in HDFS

As illustrated above, no personal customer information will be stored by the SSRM. The information collected should be stipulated in the disclaimer for the customer's knowledge.

2) Trust

Individuals should be able to trust the company's collecting their data, use the data in a responsible, predictable manner and not behave in an opportunistic manner, which leads to inappropriate ways their data are being used [16]. Mistrust is created when companies partake in unauthorised monitoring, unsolicited intrusions and security of personal data [16]. A good example of unauthorised monitoring is when a website does not disclose it uses cookies to store information of the user's browsing activity. Individuals need to feel confident companies collect their data only with their consent and use it for clearly stipulated purposes [20]. Otherwise data collection would lead to individuals feeling they are being monitored while going about everyday life activities, which is too invasive.

Another untrustworthy act from companies is sending unsolicited advertisements, emails and promotional offers, which can be difficult to opt out for individuals [21]. Their personal data could have been sold and widely distributed, so unsubscribing from one company does not usually stop the individual from being inundated by spam. Spam is usually promotional communication sent to individuals via emails or advertisements, many times without individuals consenting. Hence individuals needs to feel confident companies will ensure their data's security [16], throughout the while data security life cycle [9].

For the SSRM to be a feasible solution for the client, no subscriptions or personal contact information are needed. This means the company cannot partake in any untrustworthy acts with the customer's data. The data collected are only gender and shoe size, no email address, contact number or social media information. Also the data collected for the SSRM cannot be linked to the existing POS or customer data the company has, as no personal information will be saved in the HDFS.

3) Awareness

BD analytics is a fairly new term to a lot of consumers. Hence their knowledge of the topic is limited. Awareness of BD is the concept of the individual's understanding of what BD is, their rights regarding BD and understanding who holds the data [16]. Ethical issues arise when there is lack awareness from the individual. To eliminate this, individuals need to educate themselves in general BD policies, regulations and laws that exist to protect them from unknowingly consenting to something they do not understand. For example, in South Africa we have the Protection Of Personal Information Act, 2013 (POPI act) to "*promote the protection of personal information processed by public and private bodies*" [22]. The POPI act is quite extensive to protect the personal information of the individual and hold data owners accountable for misconduct, for example, the data owner are responsible for loss of personal information, unlawful access and processing of personal information [22]. If individuals are aware of the protection the law provides, their trust levels may rise in terms of consenting to their personal information being collected.

Individuals also need to be aware of what data are being collected about them and what third parties have access to their data [23]. Terms and conditions need to be clear for a layman to understand it and not contain mostly IT professional jargon to ease the opting out process if needed.

Customers walking into the shoe retailer will not be aware they are consenting to their data being collected via the SSRM. They will also not be aware what the retailer is going to do with the data collected. For them to be aware of it, a well informative disclaimer needs to be put up. The noninvasive nature of the SSRM/ camera and the nature of the data being collected do not heed for written consent from the customer. The disclaimer should advise the customer to ask further questions should him/ her need more information. The customer is accustomed to security cameras inside the store which they understand are there for their personal protection. If the disclaimer explains the extra camera's (to measure their shoe size) purpose is to ensure they get their shoe size in store upon their visits might make them more trustworthy and won't make them feel their rights has been violated.

4) Choice

Our own choices are one of the most fundamental things about being human. Especially a customer, a customer wants to make their own choices and these choices may change a lot. BD analytics has the ability to restrict individual's choices [24]. BD analytics can profile individuals by age, behaviors, location and gender. This leads to companies categorizing individuals according to the data collected and, only includes them in communication for services and products based on their profile [16]. As a result of this, customers may face a less-than-free market and lose out on choice [25]. Many customers feel the offers on "Takealot Daily Deals" are tailored to their historic

purchasing patterns, and all clients do not get the same discounts, although this is just a suspicion there are companies that use similar marketing strategies. Companies also target certain profiles and concentrate product/ service specific marketing on them, also offering rewards and discounts until customers start buying their products [17]. This is perceived as unethical as companies use personal data to manipulate customers in buying their products/ services.

As previously mentioned the SSRM will not be collecting any personal information on customers, so it makes it impossible target specific profiles for marketing of products. The data collected is for demand forecasting for all existing and potential customers.

III. Conclusion

BD can offer many useful benefits to companies, but it also presents unprecedented challenges. The ubiquitous nature of BD makes it invaluable to companies. This same nature makes it complex to manage. Privacy and Security preservation of the data is one of the biggest challenges companies face. Traditional security mechanisms are insufficient to protect BD, mainly because of the volume, veracity, velocity and variety characteristics. BD protection strategies should not only consider what the organization is doing internally, but from a Technological, Organisational and Environmental aspect. All these areas are where BD is vulnerable to security breaches. Thus companies should analyse and invest accordingly.

Ethical considerations are just as important as security when it comes to BD. If the source of data feels their rights or privacy has been violated, or if the company has irresponsible or malicious intent, then they will not be consensual in giving up the data. Companies need to make sure their data sources are treated respectfully and their need and rights are taken into consideration. The four main ethical aspects companies need to take into account are Privacy, Trust, Awareness and Choice. These considerations ensure mutual respect and a good relationship for future data collection.

IV. References

[1] A. Seetharaman, Indu Niranjan, Varun Tandon, and A. S. Saravanan. Impact of big data on the retail industry. Corporate Ownership and Control, 14(1):506–518, 2016.

[2] Wei Fang, Xue Zhi Wen, Yu Zheng, and Ming Zhou. A Survey of Big Data Security and Privacy Preserving. IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), 34(5):544–560, 2017.

[3] S. Vijayakumar Bharathi. Prioritizing and Ranking the Big Data Information Security Risk Spectrum. Global Journal of Flexible Systems Management, 18(3):183–201, 2017

[4] Big Data in fashion: transforming the retail sector. Journal of Business Strategy, 2019.

[5] Emmanuel Sirimal Silva, Hossein Hassani, Dag Øivind Madsen, and Liz Gee. Googling fashion: forecasting fashion consumer behaviour using google trends. Social Sciences, 8(4):111, 2019.

[6] Technological, Organizational and Environmental Security and Privacy Issues of Big Data: A Literature Review. Procedia Computer Science, 100:19– 28, 2016.

[7] Supriya Haribhau Pawar. A study on big data security and data storageinfrastructure.International Journal, 6(7), 2016.

[8] Dhirendra Sharma, Management Program, Michael Cusumano, and Thesis Supervisor. Enterprise Information Security Management Framework [EISMF] Enterprise Information Security Management Framework [EISMF]. 2011.

[9] Agata McCormac, Dragana Calic, Marcus Butavicius, Kathryn Parsons,Tara Zwaans, and Malcolm Pattinson. A reliable measure of InformationSecurity Awareness and the identification of bias in responses.Australasian Journal of Information Systems, 21:1–12, 2017.Vaibhav Kumar Sarkania and Vinod Kumar Bhalla. International Journal of Advanced Research in. Android Internals, 3(6):143–147, 2013

[10] Tianjian Zhang. Knowledge Expiration in Security Awareness Training.Annual ADFSL Conference on Digital Forensics, Security and Law, (c):197–212, 2018.

[11] Vaibhav Kumar Sarkania and Vinod Kumar Bhalla. International Journal of Advanced Research in.Android Internals, 3(6):143–147, 2013.

[12] Shiwei Sun, Casey G. Cegielski, Lin Jia, and Dianne J. Hall. Understandingthe Factors Affecting the Organizational Adoption of

[13] Big Data.Journal ofComputer Information Systems, 58(3):193–203, 2018.

[14] K. Hayashi. Social issues of big data and cloud: Privacy, confidentiality, andpublic utility. In2013 International Conference on Availability, Reliabilityand Security, pages 506–511, 2013.

[15] Richard Baskerville, Frantz Rowe, and Fran ̧cois Charles Wolff. Integrationof information systems and cybersecurity countermeasures: An exposure torisk perspective.Data Base for Advances in Information Systems, 49(1):33–52, 2018.

[16] Michael R Galbreth and Mikhael Shor. The impact of malicious agents onthe enterprise software industry.Mis Quarterly, pages 595–612, 2010.

[17] Ida Someh, Michael Davern, Christoph F. Breidbach, and Graeme Shanks.Ethical issues in big data analytics: A stakeholder perspective.Communi-cations of the Association for Information Systems, 44(1):718–747, 2019.

[18] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of aninformation civilization.Journal of Information Technology, 30(1):75–89,2015.

[19] Marcus R Wigan and Roger Clarke. Big data's big unintended consequences.Computer, 46(6):46–53, 2013.

[20] France Belanger and Robert E Crossler. Privacy in the digital age: a re-view of information privacy research in information systems.MIS quarterly,35(4):1017–1042, 2011.

[21] Solon Barocas and Helen Nissenbaum. Big data's end run around proceduralprivacy protections.Communications of the ACM, 57(11):31–33, 2014.

[22] Alexander Halavais.Bigger sociological imaginations: Framing big so-cial data theory and methods.Information, Communication & Society,18(5):583–594, 2015.

[23] https://www.justice.gov.za/inforeg/docs/InfoRegSA-POPIA-act2013-004.pdf

[24] Kate Crawford and Jason Schultz. Big data and due process: Toward aframework to redress predictive privacy harms.BCL Rev., 55:93, 2014.

[25] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of aninformation civilization.Journal of Information Technology, 30(1):75–89,2015.

[26] Mike Ananny. Toward an ethics of algorithms: Convening, observation,probability, and timeliness.Science, Technology, & Human Values, 41(1):93–117, 2016.