Security risk assessment of Machine-Learning Systems.

Amit Kumar Singh
The University of British Columbia
amitkumar.singh@alumni.ubc.ca

Mohakta Sahni
The University of British Columbia
mohakta.sahni1@gmail.com

***Abstract*— This document aims to identify top three cybersecurity risks associated with Machine Learning [1] systems. In the context of cybersecurity, Risk is defined as follows:** $Risk = Threat * Vulnerability * Consequence$ **[3] Document outline:**

1) **First part of the document will summarize the most recent work on the topic.**
2) **Second part will identify three primary Risks associated with Machine Learning systems. The Risks will be described in a table with columns presenting the constituent Consequence , Vulnerability and Threat.**
3) **Third part will summarize risk mitigation strategies discussed in academic literature from reputed publications and journals.**

I. INTRODUCTION

In the socioeconomic structure of the current age, **ML** is ubiquitous component in systems that are built with Information Technology. The popularity of **ML**, attracts researchers from diverse academic fields. Identification and categorization of top three risks associated with **ML** will be of fundamental value to the researchers who are interested in **ML**.

II. BACKGROUND

A. Problem Description

The overarching problem targeted in this work is the security and privacy vulnerabilities inherent in Machine Learning (ML) systems. Our focus is on identification of the top three risks of attack on **ML** systems.

In this section we will describe ”Consequence”, ”Vulnerability” and ”Threat” specific to the technology domain of Machine Learning.

To establish a context of the application areas pertaining to prevalent use of **ML**, we have referred to the documents mentioned in Appendix A.

B. Adversary Model

We model adversary in terms of threat and the corresponding threat agent.

III. RELATED WORK

The most recent work that is related to the security assessment of machine learning systems is presented in 2018 [1]. Appendix B summarizes this work.

IV. METHODOLOGY

The methodology deployed in this work is based on a categorization theory [4]. The visual representation of risk will be in the form of three dimensional layers.

The depth of layers will be plotted along Z dimension, which represents the ”Consequence” in our Risk assessment model. A point in risk diagram will have 0 in it’s z co-ordinate, if customer is farthest from the risk. X axis will represent ”Vulnerability” and Y axis will represent ”Threat”.

This survey instigates papers related to various threat models for different machine learning algorithms that were published in top conferences on cybersecurity and machine learning.

V. RESULTS

The result of the survey of potential risks associated with machine learning systems is presented as follows:

[1] ”Machine Learning” is abbreviated as**ML** in the rest of the document

Risk Factors	Threat	Vulnerability	Consequence
Domain			
Healthcare	NaN	NaN	NaN
Finance	NaN	NaN	NaN
Governance	NaN	NaN	NaN
Defense	NaN	NaN	NaN
Environment	NaN	NaN	NaN
Society	NaN	NaN	NaN

VI. Discussion

VII. Conclusion

VIII. Appendices

A. Appendix A

This section of the document summarizes the literature that were referenced to establish the context of Machine Learning systems. [2]

1) Machine Learning is described as a technology that address the need of automated data analysis.
   Target audience, as described in the book -

   ”This book is suitable for upper-level undergraduate students and beginning graduate students in computer science, statistics, electrical engineering, econometrics, or any one else who has the appropriate mathematical background. Specifically, the reader is assumed to already be familiar with basic multivariate calculus, probability, linear algebra, and computer programming. Prior exposure to statistics is helpful but not necessary.”
2) (ML) is categorized in three major areas, viz. ”Supervised Learning”, ”Unsupervised Learning” and ”Reinformcement Learning”. Vulnerability of an ML system is seen as an information flow pipeline that begins with input features, digital representation of input features, learning mechanism to learn from input features, deployment of learned system. [1]

B. Appendix B

The SoK [1], presents valuable information on ”attack” and ”defenses” as applicable to machine learning systems. The following table puts the summary of the paper in perspective.

On Vulnerability	Attack	Defense
ML Category		
Supervised	NaN	NaN
Unsupervised	NaN	NaN
Reinforcement	NaN	NaN

IX. References

References

[1] Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018, April). SoK: Security and privacy in machine learning. In 2018 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 399-414). IEEE.

[2] Robert, C. (2014). Machine learning, a probabilistic perspective.

[3] Cox, Jr, L. A. (2008). Some limitations of Risk= Threat Vulnerability Consequence for risk analysis of terrorist attacks. Risk Analysis: An International Journal, 28(6), 1749-1761.

[4] Cisco, S. L., & Jackson, W. K. (2005). Creating order out of chaos with taxonomies. Information Management, 39(3), 44.

[5] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu and V. C. M. Leung, ”A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View,” in IEEE Access, vol. 6, pp. 12103-12117, 2018. doi: 10.1109/ACCESS.2018.2805680

[6] Battista Biggio, Fabio Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recognition, Volume 84, 2018, Pages 317-331, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2018.07.023.

[7] B. Nelson, M. Barreno, F.J. Chi, A.D. Joseph, B.I.P. Rubinstein, U. Saini, C. Sutton, J.D. Tygar, K. Xia, Exploiting machine learning to subvert your spam filter Proceedings of the LEET, USENIX Association (2008), pp. 1-9

[8] B. Biggio, I. Corona, B. Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, F. Roli Security evaluation of support vector machines in adversarial environments Y. Ma, G. Guo (Eds.), Support Vector Machines Applications, Springer International Publishing, Cham (2014), pp. 105-153

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. rndi, P. Laskov, G. Giacinto, F. Roli Evasion attacks against machine learning at test time Proceedings of the ECML PKDD, Part III, LNCS, 8190, Springer (2013), pp. 387-402

[10] M. Barreno, B. Nelson, A. Joseph, J. Tygar The security of machine learning Mach. Learn., 81 (2010), pp. 121-148

[11] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D.Verma Adversarial classification Proceedings of the ICKDDM (2004), pp. 99-108

[12] Jianfeng Wen, Shixian Li, Zhiyong Lin, Yong Hu, Changqin Huang, Systematic literature review of machine learning based software development effort estimation models, Information and Software Technology, Volume 54, Issue 1, 2012, Pages 41-59, ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2011.09.002.

[13] N. Papernot, P. McDaniel, A. Sinha and M. P. Wellman, ”SoK: Security and Privacy in Machine Learning,” 2018 IEEE European Symposium on Security and Privacy (EuroS& P), London, 2018, pp. 399-414. doi: 10.1109/EuroSP.2018.00035

[14] Anshuman Singh, Sumi Singh, Andrew Walenstein, and Arun Lakhotia. 2011. On deployable adversarial classification models. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (AISec ’11). ACM, New York, NY, USA, 113-114.